# Librarian Center

- Librarian Center Home
- **Newsletter Archive**
- Downloads
- Contact Us

## How does Google collect and rank results?

**One of the most common questions we hear from librarians is "How does Google decide what result goes at the top of the list?" Here, from quality engineer Matt Cutts, is a quick primer on how we crawl and index the web and then rank search results. Matt also suggests exercises school librarians can do to help students**.

**Crawling and Indexing**
A lot of things have to happen before you see a web page containing your Google search results. Our first step is to crawl and index the billions of pages of the World Wide Web. This job is performed by Googlebot, our "spider," which connects to web servers around the world to fetch documents. The crawling program doesn't really roam the web; it instead asks a web server to return a specified web page, then scans that web page for hyperlinks, which provide new documents that are fetched the same way. Our spider gives each retrieved page a number so it can refer to the pages it fetched.

Our crawl has produces an enormous set of documents, but these documents aren't searchable yet. Without an index, if you wanted to find a term like *civil war,* our servers would have to read the complete text of every document every time you searched.

So the next step is to build an index. To do this, we "invert" the crawl data; instead of having to scan for each word in every document, we juggle our data in order to list every document that contains a certain word. For example, the word "civil" might occur in documents 3, 8, 22, 56, 68, and 92, while the word "war" might occur in documents 2, 8, 15, 22, 68, and 77.

Once we've built our index, we're ready to rank documents and determine how relevant they are. Suppose someone comes to Google and types in *civil war*. In order to present and score the results, we need to do two things:

1. Find the set of pages that contain the user's query

somewhere

2. Rank the matching pages in order of relevance

We've developed an interesting trick that speeds up the first step: instead of storing the entire index on one very powerful computer, Google uses hundreds of computers to do the job. Because the task is divided among many machines, the answer can be found much faster. To illustrate, let's suppose an index for a book was 30 pages long. If one person had to search for several pieces of information in the index, it would take at least several seconds for each search. But what if you gave each page of the index to a different person? Thirty people could search their portions of the index much more quickly than one person could search the entire index alone. Similarly, Google splits its data between many machines to find matching documents faster.

How do we find pages that contain the user's query? Let's return to our civil war example. The word "civil" was in documents 3, 8, 22, 56, 68, and 92; the word "war" was in documents 2, 8, 15, 22, 68, and 77. Let's write the documents across the page and look for those with both words.

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| civil | 3 | 8 | | 22 | 56 | 68 | 92 |
| war | 2 | 8 | 15 | 22 | | 68 | 77 |
| both words | | 8 | | 22 | | 68 | |

Arranging the documents this way makes clear that the words "civil" and "war" appear in three documents (8, 22, and 68). The list of documents that contain a word is called a "posting list," and looking for documents with both words is called "intersecting a posting list." (A fast way to intersect two posting lists is to walk down both at the same time. If one list skips from 22 to 68, you can skip ahead to document 68 on the other list as well.)

### An exercise for students

Once you see how to intersect two words in an index, it's not hard to do it for three or more words as well. Here's a fun exercise: try to find all the documents below that contain the words "civil" and "war" and "reconstruction."

civil: 1 9 15 19 22 35 38 48 53 55 65 68 73 78 82 88 91 99
war: 15 18 25 29 31 35 37 40 42 46 48 65 75 85 91 96
reconstruction: 35 42 48 64 73 91 95

The answer is at the end of the article.

### Ranking Results

Now we have the set of pages that contain the user's query somewhere, and it's time to rank them in terms of relevance. Google uses many factors in ranking. Of these, the PageRank algorithm might be the best known. PageRank evaluates two things: how many links there are to a web page from other pages, and the quality of the linking sites. With PageRank, five or six high-quality links from websites such as www.cnn.com and www.nytimes.com would be valued much more highly than twice as many links from less reputable or established sites.

But we use many factors besides PageRank. For example, if a document contains the words "civil" and "war" right next to each other, it might be more relevant than a document discussing the Revolutionary War that happens to use the word "civil" somewhere else on the page. Also, if a page includes the words "civil war" in its title, that's a hint that it might be more relevant than a document with the title "19th Century American Clothing." In the same way, if the words "civil war" appear several times throughout the page, that page is more likely to be about the civil war than if the words only appear once.

**An exercise for students**

Pretend that you're a search engine. Pick a query like *civil war* or *recycling* or whatever you want. Search for the phrase on Google, pick three or four pages from the results, and print them out. On each printout, find the individual words from your query (such as "civil" and "war") and use a highlighter to mark each word with color. Do that for each of the 3-5 documents that you print out. Now tape those documents on a wall, step back a few feet, and squint your eyes. If you didn't know what the rest of a page said, and could only judge by the colored words, which document do you think would be most relevant? Is there anything that would make a document look more relevant to you? Is it better to have the words be in a large heading or to occur several times in a smaller font? Do you prefer it if the words are at the top or the bottom of the page? How often do the words need to appear? See if you can come up with 2-3 things you would look for to see if a document matched a query well. This can help students learn to evaluate website relevance the way a search engine would evaluate it so that they can better understand why a search engine returns certain results over others.

As a rule, Google tries to find pages that are both reputable and relevant. If two pages appear to have roughly the same amount of information matching a given query, we'll usually try to pick the page that more trusted websites have chosen to link to. Still, we'll often elevate a page with fewer links or lower PageRank if other signals suggest that the page is more relevant. For example, a web page dedicated entirely to the civil war is often more useful than an article that mentions the civil war in passing, even if the article is part of a reputable site such as Time.com.

Once we've made a list of documents and their scores, we take the documents with the highest scores as the best matches. Google does a little bit of extra work to try to show snippets – a few sentences – from each document that highlight the words that a user typed. Then we return the ranked URLs and the snippets to the user as results pages.

As you can see, running a search engine takes a lot of computing resources. For each search that someone types in, over 500 computers may work together to find the best documents, and it all happens in under half a second.

Did you know? On April 1, 2002, we spoofed our PageRank algorithm by presenting a detailed explanation of "PigeonRank"

[Answer: Only documents 35, 48, and 91 contain all three words "civil" and "war" and "reconstruction."]

Matt Cutts is a software engineer in the quality group at Google. He spends his days trying to help good sites rank where they should and developing techniques that keep deceptive or spammy sites from showing up in Google's search. He also has a web log at http://www.mattcutts.com/blog/ that often discusses webmaster issues.

Other questions? Send us a note. Every newsletter we'll try to answer 1 or 2 of the most frequently asked questions.

Sign up to receive this newsletter.