Search Engines

- Browser uses a FORM to send a query to a server
  - e.g., google.com
- Server runs a program to extract query from form
- Finds pages that contain word(s) of query
- Generates HTML
- Returns page to client
- Server needs to know what pages contain relevant words
- Continuously crawls the web collecting pages
- Builds big database that tells what pages contain any given word
- Basic problem: scale
  - lots of pages, lots of words, lots of queries

Server Processes

- 3 basic processes going on in parallel
  - respond to incoming queries by looking up words in database
  - crawl web looking for new pages
  - extract words from new pages and insert into database

Fetching New Pages

- Start with a list of likely URLs
- Fetch data from next URL from the list
  - obey robot exclusion standard
- Extract parts to be indexed, deliver to index builder
- Extract URLs
- Delete duplicate URLs (ones seen recently)
• Delete irrelevant ones (advertisements, ...)
• Add remaining URLs to end of list
• Go back to the top
• Questions:
  – How to start?
  – How to detect duplicates quickly?
  – What to preserve (text, .html files, .txt files, PDF, gif/jpg, ...)?
  – How to avoid overloading big/popular sites?

Building and Searching an Index

• For a new page that has just been fetched:
  – isolate words (discard HTML tags, etc.)
  – handle upper and lower case, accents, punctuation, other languages and character sets, ...
  – for each word
    add URL to list for that word
    add word position within the page to the list for the URL
• To look up a single word query:
  – go to the list for the word
  – collect all URLs
  – sort them into order by weighting function
    importance, frequency, ...
• Queries with multiple words:
  – collect URL lists, combine them, weight them

Hashing: An Algorithm to Look Things Up Quickly

• Problem: how to look up one word in 1 billion words, really fast
  – binary search would be 30 probes if names were sorted
  – sorting takes too long if it has to be updated
• Hashing: scramble the word into an integer
  – between 0 and N
  – so that hash values of potential words are spread out uniformly
• Store all words with the same hash value together
• Searching for a word then requires only
  – compute the hash value
  – look at the list of previously-stored words with that hash value
• Example hashing algorithm: add up the numeric values of all the characters in the word

Ranking Search Results

• How to get the most likely results on the first page (at the top)
  – most people look only at the first few results
  – need for very high precision (relevant documents in the top 10 or so)
• Google uses proprietary "page rank" algorithm based on link structure of web
  – pages that are cited often move higher
- pages that are cited by higher ranked sites move higher
- anchor (<a href=...>) text gives more information
- proximity of search terms within page
- ...

• Other search engines have analogous techniques
• Have to defend against attempts to inflate rankings

Privacy and Copyright Issues

• What privacy standards apply to search engines?
  – how can private / incorrect information be purged?
• Search engines versus government
  – should search engines release information about dissidents to the local government?
  – should search engines suppress / restrict query results if requested by government?
  – can query logs be subpoenaed?
    - AOL's release of "sanitized" information permitted identification of individuals from their queries
• Copyright
  – Viacom v YouTube: vicarious liability or DMCA safe harbor?
  – should newspaper stories be indexed without permission?
• Trademarks
  – can someone buy someone else's trademark as an advertising keyword?
    - e.g., Microsoft buys "iPod"
• ...

More about Google and Search

Tim Berners-Lee's dream has become reality. We are adding millions of web pages per day about anything and everything. Now the problem is finding the information that interests us. There is no table of contents or index to search. This lack of navigation was recognized early. Yahoo had tried to use user input to create a directory tree but was too cumbersome. The first search engine, Lycos, was created in 1994 at Carnegie Mellon University.

As seen, search engines use databases to record the words found inside web pages. They rely on two separate, automated programs to track web pages and the words within them:

1. **Spiders** (robots, crawler) – Software that methodically trawl the Web by downloading copies of pages as they traverse web sites.
2. **Indexer** – Records the text on the downloaded pages along with important information that is encoded within them, i.e.,
   - page title
   - links to other pages

All of this information is stored into a database called an "index" or "catalog." It is kept up-to-date by running the spiders often (usually early in the morning when Internet traffic is light.) Conclusion: search engines search their index not the WWW (which would be impossible). Google has 16 billion web pages indexed as of 2014.
Early search engines used text counting to rank pages – the more often a term appeared, the higher the page was in the result list. Once web site owners understood this, they began to increase terms within pages to increase their rank; called **site manipulation**. Site manipulation led to unsatisfactory search results.

In 1995, **Sergey Brin** and **Larry Page** at Stanford University created “**PageRank**.” This new idea did not rank by page content but by community input – a site is ranked primarily on how many other sites link to it and ranked search results accordingly. Now results tended to be more useful and relevant.

In 1998, Google had four employees working out of a garage, whose server was handling 9,800 search queries per day. Today, they employ 900,000 servers to handle 5,740,000,000 search queries per day. Also, Google has 53,600 employees. A share of Google stock sold for $54.21 a share at its initial offering, today it sales for between $562.30 and $575.52 a share. Search was important for the WWW!

**How PageRank works:**

1. Examine the links within a web page, award votes for second page status, e.g., MLB.com found on many baseball related web sites.
2. Voting is weighted by the popularity of the page voting.
3. PageRank returns a value of 1 to 10 to rank status.
4. Now combine PageRank score and text matching, i.e., the content of matching pages and the pages linked to them

One major strength of Google's approach was that they separated paid sites from search sites.

**Google User Interface**

Simple UI to suggest speed and search only focus. Some changes over time, logo now has animations and “I'm Feeling Lucky” has been disabled. (Showed only the top ranked page found.)

**Search Fundamentals**

Not easy, consider the word “apple.” It is a fruit, a computer company, the Beatles' record label, a nickname for New York City, a singer named Fiona Apple, etc.. But the computer company would rise to the top because there exists more links to Apple Computer than to the other choices.

The rule is Garbage In, Garbage Out (GIGO), so search terms are important for quality results. To obtain information about the other apples, you need to add key words, e.g.,

apple nutrition Apple Records Big Apple Fiona Apple
Search Tips and Techniques

1. Don't use questions – use answers instead. Instead of "how many calories does an apple have" use "an apple has * calories" instead. [* is a used to imply a single word]
   a. Singular is different than plural.
   b. The ordering of words matters – the first word is the most important term.
   c. Google ignores most “little words” (or “stop words”). Examples of ignored words:
      where how the of an for from it in
   d. Use quotation marks if you want to include a small word, e.g., “The King”
   e. Most punctuation is also ignored except apostrophes, hyphens, and quotation
      marks.
   f. Use .. to express a range of numbers. 1 .. 10

2. Use quotation marks around whole phrases if the entire phrase is significant.

3. AND vs OR – the default behavior is to search for all words listed. If you insert AND
   between all the search words the resulting pages must include all the words. If you
   insert OR between the words you get any one of your search words. Use parentheses
   for grouping, e.g.
   
   chimps (“Ben Affleck" OR “Mark Wahlberg")
   Note: | can be used instead of OR

4. NOT is not understood, instead use – (minus sign)
   “Ben Affleck” -“Jennifer Lopez”
   Note that minus sign must appear directly before the word or phrase but you need a
   space in front of the minus sign.

5. + is used to include but quotes have the same results.

6. Use the tabs. On the top of every search are a number of tabs. Usually you’ll see Web,
   Image, News, and More. Using these tabs, you can help define what kind of search
   you need to do. They can cut search times dramatically if utilized properly.

Google Quirks

1. Wildcards are not allowed but a single * is used for “full-word” substitution. Multiple
   asterisks will increase the number of words to substitute. “Chicken with * cut off"
2. Ten word limit per search but doesn't count asterisks within search.

Cool Google Tricks

1. **Start with the basics.** No matter what you're looking for, start with a simple search. If
   you're looking for a place or product in a specific location, add the location.
2. **Choose words carefully.** When you're deciding what words to put in the search box,
   try to choose words that are likely to appear on the site you're looking for.
3. **Don’t worry about the little things.** Spelling: Google's spell checker automatically
   uses the most common spelling of a given word, whether or not you spell it correctly.
   Capitalization: search is case insensitive.
4. **Find quick answers.** For many searches, Google will do the work for you and show
   an answer to your question in the search results.
   weather location
   define word
   Enter any math equation
   Enter any conversion, like 3 dollars in euros.