Search & Google

Melissa Winstanley
mwinst@cs.washington.edu

The size of data

- Byte: a single character
- Kilobyte: a short story, a simple web html file
- Megabyte: a photo, a short song
- Gigabyte: a movie, a pickup truck filled with paper covered in text
- Terabyte: 50,000 trees worth of printed paper, all the x-ray films for a hospital
- Petabyte: half of all of the US academic research libraries
- Exabyte: total mobile traffic per month
- Zettabyte: total digital information
  - 1,000,000,000,000,000,000,000 bytes

Some history

- Archie (1990): indexed files but only examined the file names
- WebCrawler (1994): Indexed the entirety of documents
- AltaVista (1995): Indexed a good portion of the Internet
- Lycos (1995): incorporated links in their algorithms
- Yahoo (1995): Worked as a directory, not a search engine
- Google (1998): Invents PageRank and has dominated the search industry ever since

Google Map Cars

So how does Google do it?

- A zettabyte of data is a lot
- We probably can't search every document for your query
    - Would take a long time!
- Need to be cleverer

Another challenge:

- Most of digital data is video and photos
- How do you search these?

Document index

- We can think of a document as words
- One thing we can do is create a mapping from documents to words (no duplicates)
  - Notice that these are sets of words

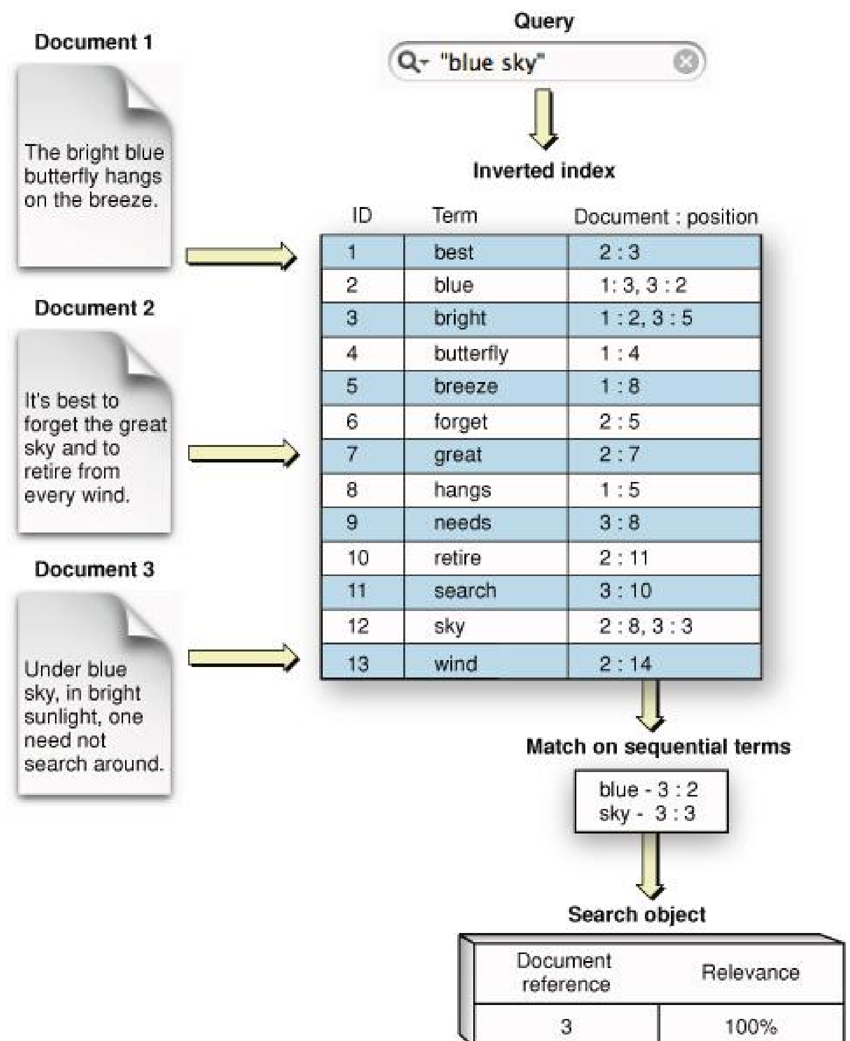
Document	Words
1	The, quick, brown, fox, jumped, over, the, lazy, dog
2	The, cow, jumped, over, the, moon
3	Crazy, like, a, fox
4	The, man, in, moon

Searching the index

- Given a query Q
  - Look through each document and see if it's there
- What's the problem with this?
  - **O(N)**, where N is the total number of documents
  - Most documents don't contain Q
- What we really want to do is get a mapping in the other direction
  - From words/queries to documents
  - Then lookup would be fast

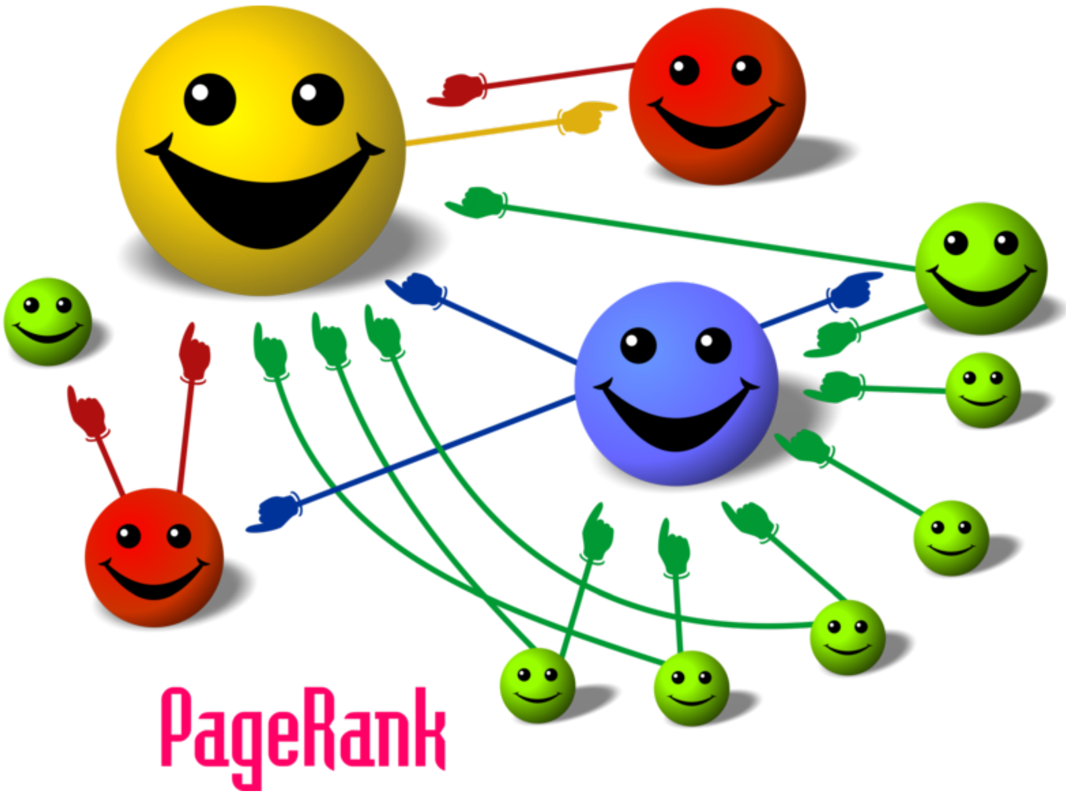
Inverted Index

- Catalog what words appear in each document
  - Words --> sets of documents that contain the words
- Query --> returns documents a word appears in
- Phrase searches --> documents that appear in all the words' lists



Pagerank

- PageRank is a way to rank websites by importance
- Pretend that a user randomly clicks links on the web forever
- % of time spent on a given site



Pagerank

$$\underbrace{PR(A)}_{\text{Pagerank of page A}} = \underset{\uparrow\atop\text{Sum for all pages that link to A}}{\sum} [\underbrace{PR(i)}_{\text{Pagerank of page i}} / \underbrace{L(i)}_{\text{Number of links out of page i}}]$$

Rank for a page is the probability of getting there from every other page, weighted by their pageranks.

Improving searches

- Get rid of capitalization and punctuation
  - "Hello!" --> "hello"
- Blacklist common, useless words
  - the, and, a, etc.
- Stemming
  - "swimming" --> "swim"
- Ngram indexing
  - "red balloon" not just "red" and "balloon"

More improvements

- Context
  - Words that occur near the search that are related
- Spell check
  - Look for words spelled similarly, but are more
- Frequency analysis
  - Words that appear more are likely important
  - BUT - files of different lengths
- Within-file location
  - Then you can search around a term
- Page link text, titles, headers

"Search Engine Optimization"

- What could we do to boost our search ranking for a particular term?
  - Increase our page rank
  - Include the number of search term occurrences
  - Buy links from high ranking web sites
- Responses by Google and other search engines
  - No-follow ref attribute on links
  - robots.txt

Accuracy

- Google maps is so popular that their content is used all over the world
- The data is often held as authoritative, which in some cases is not a good thing
  - Morocco and Spain Land Dispute
  - Nicaragua and Costa Rica Land Dispute
- When you have so much data being accessed by so many people accuracy becomes a huge issue

Google flu trends

- Millions of searches are made through Google a day
  - We can gather many statistics and facts about the world by sifting through this information
- In this case Google found that search terms for the flu were actually indicative of the flu
- In particular, Google's data seems to be about 2 weeks ahead of the CDC
- Google hosts data for many countries and is even experimenting with cities

Google Insights & Trends

- Insights is a web app that lets us look up metasearch data
  - How frequently is a search query made?
  - Where are they made?
  - How has its popularity changed over time?
  - What related searches are made?
- Trends tells us what search terms are the most popular right now

New problems

- Images
  - How does one search these?
  - Identify *features* - lines, textures, colors
  - Mathematical model of image
  - Match features with a database
- Video
  - Same problems as photos, but more
  - Audio feature analysis
  - Still frame analysis
  - Video-specific features

New problems

- Too much data
  - More data than can fit on one computer
  - How do we store data on multiple computers?
  - "Distributed systems"
  - If interested, look at the Google File System
  - (see operating systems course - CSE 451)

Now what?

- We have a way to search all this data
- What could we do with it?

- "Knowledge graph"
  - Building up a foundation of knowledge
  - Add *meaning* to a search