

Exploration Session 5: Searching and Google

Melissa Winstanley
mwinst@cs.washington.edu

(based on slides by Roy McElmurry)

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

THE HISTORY OF SEARCH

- 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



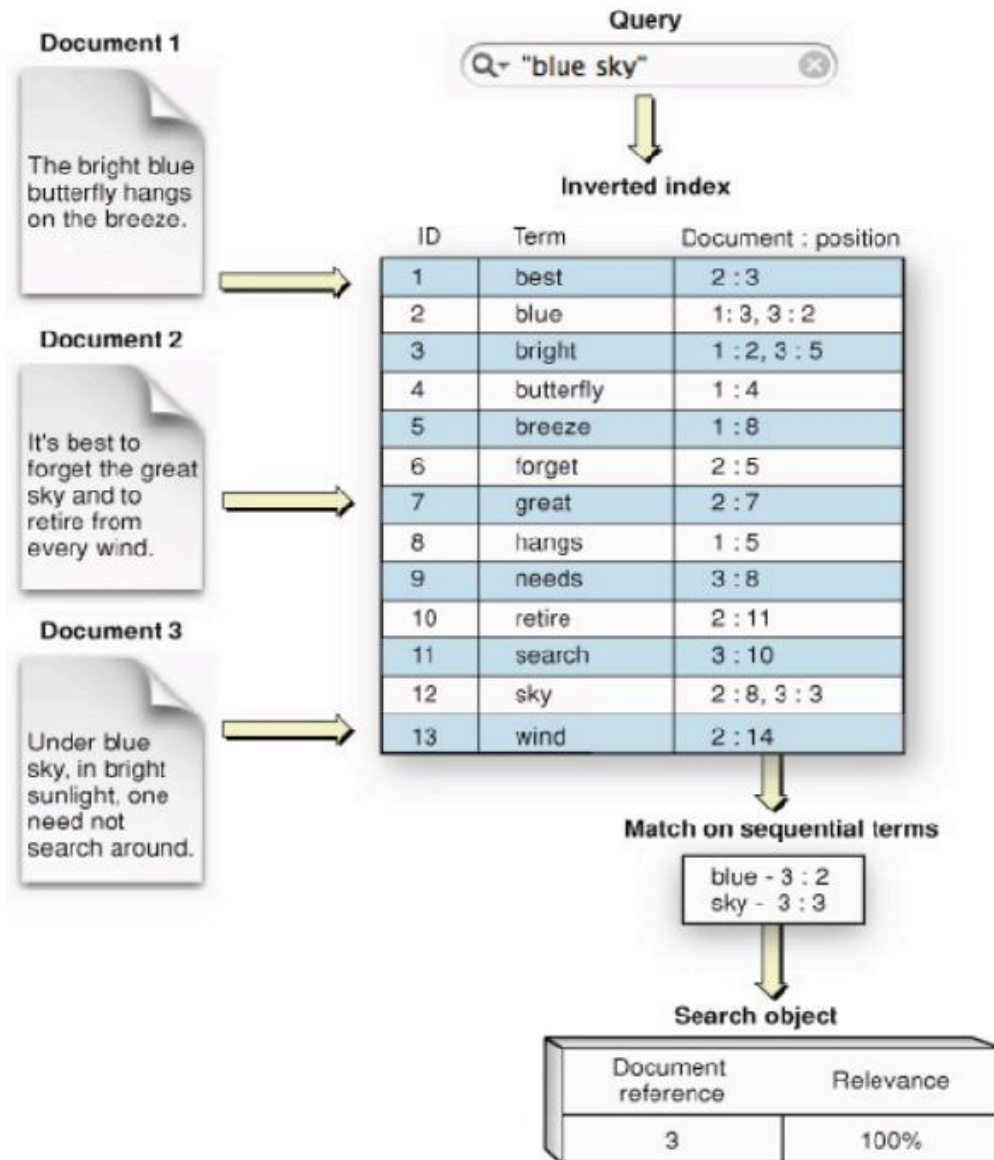
DOCUMENT INDEX

- We can think of a document as a set of words
- One thing we can do is create a mapping from documents to sets of words
 - Notice that these are sets of words – no duplicates
- What we really want to do is get a mapping in the other direction

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

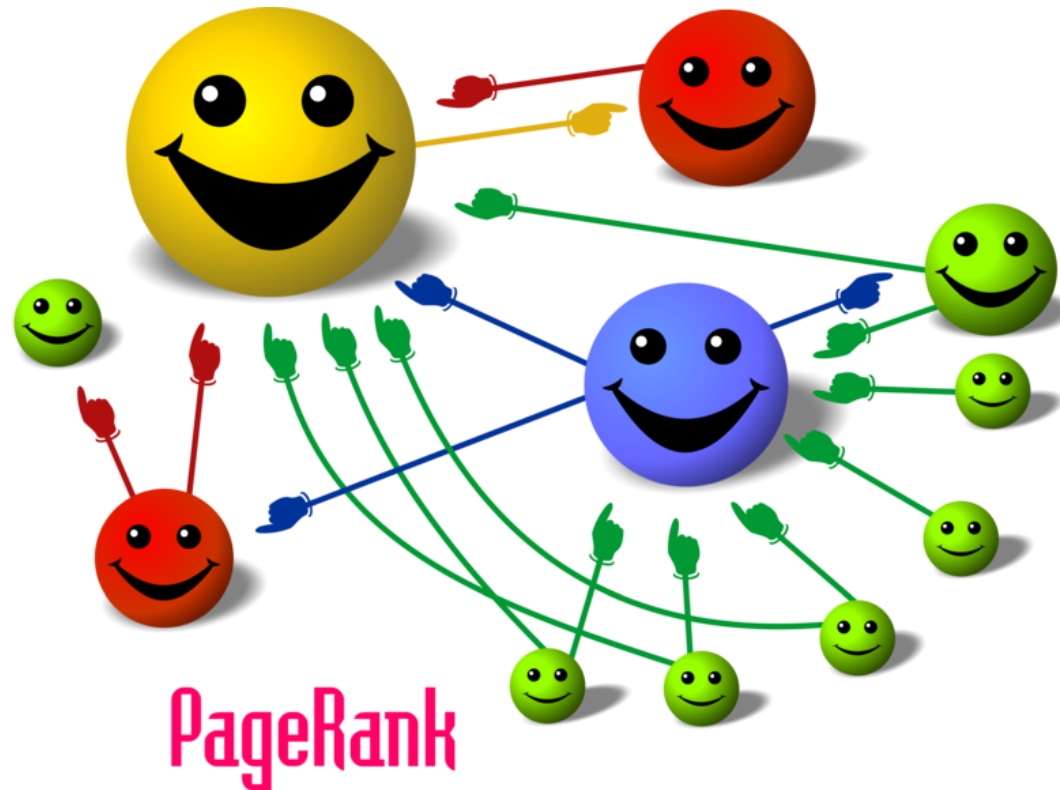
INVERTED INDEX

- Catalog what words appear in each document
 - Mapping from words to sets of documents that contain the words
- When we search for a term we can then go backwards and retrieve the documents that contain the search phrase
- To do phrase searches we simply take the intersection of the returned sets for each component word



PAGERANK

- PageRank is a way to rank websites by importance
- Pretend that a user randomly clicks links on the web forever
- PageRank is intended to determine the percentage of time one would spend on a given site following the above model
- $PR(A) = \sum PR(i) / L(i)$ where $L(i)$ is the number of links coming out of page i



SEARCH ENGINE IMPROVEMENTS

- Stop words / blacklist
- Term frequency / inverse
- Stemming
- Ngram indexing
- Parsing out punctuation
- Context analysis

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

GOOGLE MAPS DATA ERROR

- 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
- Site is [here](#)
- 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 AND TRENDS

- Insights is a web app that lets us look up meta-search 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