How does Google search for everything? Searching For and Organizing Data

Prof. Susan Rodger
Computer Science Dept
Duke University
Oct. 29, 2010

Lots of names. How do we find someone?

- Anderson, Mary : 203 Main St. Durham NC
- Williams, Fred : 14 Union Circle, Cary, NC
- Wu, Xin : 57 Wilson Court, Raleigh, NC
- Smith, Xin : 57 Wilson Court, Raleigh, NC
- Pratt, Sarah : 6 White Lane, Hillsborough, NC
- Chase, Angela : 34 Dogwood Road, Durham, NC
- Brooks, Bolton : 10 Time St., Durham, NC
- French, Bolton : 10 Time St., Durham, NC
- Gao, Bo : 134 Brookside Lane, Durham, NC

Put the names in alphabetical order

- Anderson, Mary : 203 Main St. Durham NC
- Brooks, Bolton : 10 Time St., Durham, NC
- Chase, Angela : 34 Dogwood Road, Durham, NC
- French, Bolton : 10 Time St., Durham, NC
- Gao, Bo : 134 Brookside Lane, Durham, NC
- Pratt, Sarah : 6 White Lane, Hillsborough, NC
- Smith, Doug : 18 Pine Cone Lane, Durham, NC
- Williams, Fred : 14 Union Circle, Cary, NC
- Wu, Xin : 57 Wilson Court, Raleigh, NC

Find Narten

- Narten
- Munson
- Find
- How many words did we look at?
Searching for words

• If we had a million words in alphabetical order, how many would we need to look at worst case to find a word?

• About 20

How does one search for an item?

• Data must be organized in some way
• Sorting alphabetically (or numerically) is one way
• There are other ways to organize data!

Google Search Query

1. The web server sends the query to the index servers. The content inside the index servers is similar to the index in the back of a book - it tells which pages contain the words that match the query.

2. The query travels to the document servers, which actually retrieve the stored documents. Snippets are generated to describe each search result.

3. The search results are returned to the user in a fraction of a second.

Computer Science at work behind the scenes!

• Googlebot web crawler
  – Finds and retrieves pages
  – Gives pages to google indexer
Google is all about problem solving and writing algorithms

- Algorithms must happen fast!
- Can Google put all the web pages it finds in alphabetical order to search?
- Want efficient, fast algorithms!
- No one wants to wait on a search query!
Activities

• Given numbers – sort yourselves
• Redistribute numbers – sort using selection sort
• Parallel Sort
• Hashing with buckets
  – Hash function is last digit – remainder when you divide by 10

Sort numbers (largest at bottom) using comparators in parallel

Sorting Network
different setup for comparators
Sort numbers (largest at bottom)

My research - Making theoretical concepts come alive – visualize and interact with!