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

Prof. Susan Rodger
Computer Science Dept
Duke University
Oct. 31, 2014
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, Doug : 18 Pine Cone Lane, Durham, NC
- Pratt, Sarah: 6 White Lane, Hillsborough, NC
- Chase, Angela: 34 Dogwood Road, Durham, NC
- Brooks, Bolton : 10 Time St., Durham, NC
- French, Melvin : 42 Starship Circle, 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, Melvin : 42 Starship Circle, 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

Found!

How many words did we look at?

Narten

Found!
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?

<table>
<thead>
<tr>
<th>Number of Words</th>
<th>1,000,000</th>
<th>50,000</th>
<th>25,000</th>
<th>12,500</th>
<th>6,250</th>
<th>3,125</th>
<th>1,562</th>
<th>781</th>
<th>390</th>
</tr>
</thead>
<tbody>
<tr>
<td>Worst Case</td>
<td>195</td>
<td>97</td>
<td>48</td>
<td>24</td>
<td>12</td>
<td>6</td>
<td>3</td>
<td>2</td>
<td>1</td>
</tr>
</tbody>
</table>

If you are clever,
Cut the number of numbers to look at in half over and over again,
Then only 18 numbers to look at worst case
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 doc 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
how google search works

How Google Works - Google Guide
Feb 2, 2007 ... Google gives more priority to pages that have search terms near each other ... For more information on how Google works, take a look at the ...
www.googleguide.com/google_works.html - Cached - Similar

Google Technology
The heart of Google’s search technology is PigeonRank™, a system for ranking web pages ... Building upon the breakthrough work of B. F. Skinner, ...
www.google.com/technology/pigeonrank.html - Cached - Similar

Corporate Information - Technology Overview
The software behind our search technology conducts a series of simultaneous ... The life span of a Google query normally lasts less than half a second, ...
www.google.com/corporate/tech.html - Cached - Similar

HowStuffWorks "The Google Search Engine"
The Google search engine allows you to find just about anything you can think of on the Internet. Learn more about the Google search engine.
computer.howstuffworks.com/google1.htm - Cached - Similar

HowStuffWorks "How Internet Search Engines Work"
Internet search engines do your research for you. Learn how internet search engines like Google work, how internet search engines build an index and what ...
computer.howstuffworks.com/search-engine1.htm - Cached - Similar

The Anatomy of a Search Engine
The Google search engine has two important features that help it produce high ...... That works out to be about 850 terabytes. Also assume that indexing a ...
infolab.stanford.edu/~backrub/google.html - Cached - Similar
Correction Algorithms
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
Sorting Network
Sort numbers (largest at bottom) using comparators in parallel

<table>
<thead>
<tr>
<th>8</th>
<th>4</th>
<th>4</th>
<th>4</th>
<th>4</th>
<th>2</th>
<th>2</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>8</td>
<td>6</td>
<td>6</td>
<td>2</td>
<td>4</td>
<td>1</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>8</td>
<td>2</td>
<td>6</td>
<td>1</td>
<td>4</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>2</td>
<td>8</td>
<td>1</td>
<td>6</td>
<td>3</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>7</td>
<td>1</td>
<td>8</td>
<td>3</td>
<td>6</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>7</td>
<td>3</td>
<td>8</td>
<td>5</td>
<td>6</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>3</td>
<td>3</td>
<td>7</td>
<td>5</td>
<td>8</td>
<td>7</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>7</td>
<td>7</td>
<td>8</td>
<td></td>
</tr>
</tbody>
</table>
Sorting Network

different setup for comparators
Sort numbers (largest at bottom)

<table>
<thead>
<tr>
<th>8</th>
<th>4</th>
<th>4</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>8</td>
<td>7</td>
<td>6</td>
</tr>
<tr>
<td>7</td>
<td>6</td>
<td>6</td>
<td>7</td>
</tr>
<tr>
<td>6</td>
<td>7</td>
<td>8</td>
<td>8</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>1</td>
<td>5</td>
<td>5</td>
<td>5</td>
</tr>
</tbody>
</table>
My research - Making theoretical concepts come alive – visualize and interact with!