EECS 395/495: Web Information Retrieval and Extraction

Fall 2012
Outline

• Introductions
• Course goals and logistics
• Why take this class?
• What is Web Search?
  – What will it be in five years? Ten years?
Introductions

• Professor: Doug Downey
  – Ford 3-345
Goals

• Learn:
  – How does Web search work?
  – What’s the future of Web search?

• Learn how to:
  – Analyze, discuss, and present research papers
  – Do projects at the frontier of Web search and information extraction
Logistics

- Tuesdays: lectures

- Thursdays: debates on research papers
  - Offense and Defense (2-3 on each side)
    - Defense (in favor of paper) starts with a 10 minute presentation on the key ideas
      - Then, offense/defense debate, with participation from all
      - Debaters graded by prof and peers
    - See Web page and sign up (e-mail TA) this week
Logistics

• Grading
  – Participation (30%)
    • During lectures/discussions (5%)
    • Leading a debate (15%)
    • Synopses of papers (10%)
  – Projects (70%)
    • Project grade, based on individual contribution (60%)
    • Review of another team’s project (10%)
Paper synopses

• About two paragraphs, due each Thursday
  – Starting **Thursday, Oct 11**
  – Turn-in via Blackboard

• Include:
  – Brief summary of what the paper says and its main contributions
  – One potential flaw in the paper
  – One suggested “next step” for the paper
Projects

• Groups of 2-4 (*not necessarily = debate groups*)

• Examples:
  – Read 3-5 recent research papers and summarize the state of the art in some area of IR/IE
  – Implement an IR/IE system and report on the results
  – Answer theoretical questions
  – Etc.
Specific Examples

• Systems
  – Create a search “knob” for specifying reading level
  – Execute relevant background searches as I write an e-mail
  – Automatically monitor bias in search engines

• Theoretical questions
  – Read a paper on search ad auctions or PageRank and attempt to extend the results

• Summary of Research
  – Read 5 papers on automated question answering; summarize the field and suggest future directions
Project Milestones

• **October 16**: Proposal (~1 page)
  – Meetings with me to finalize **Oct 17-18**

  ...*lots of progress*...

• **Nov 19**: Report preliminary results (~2 pages)
  – **Nov 21**: Review group provides feedback (~1 page)
• **Dec 5**: Final Report (~4 pages)
• **Dec 6/10**: Final Presentations
  (~8 mins + 3 min for Q&A)
Outline

• Introductions
• Course goals and logistics
• **Why take this class?**
• What is Web Search?
  – What will it be in five years? Ten years?
Who cares about search?

• "The most important application for the foreseeable future...is search."
  – Steve Ballmer, CEO of Microsoft
  

• Why?
  – Search’s utility scales as the Web scales
  – People use it all the time
  – Control
  – Profit
Why should you care about search?

• Opportunity
• Fascinating & important enabling technologies:
  – Scaling
  – Machine Learning/Data Mining
  – Graph-based algorithms
  – Language Understanding
  – Auction theory
  – User Interfaces
  – …
Graph-based Algorithms

• Example: PageRank
  – Google’s original claim to fame
  – Idea: Quality of $p$ is proportional to the aggregate quality of the pages linking to $p$

• $\text{PageRank}(p) = \text{probability that a “random surfer” lands on } p$
  – Pick a starting page at random
  – Follow links uniformly at random
    • Every now and then, jump to a random page
  – $\text{PageRank}(p) = \text{proportion of visits to } p$
PageRank example

15% probability of a random jump
How to compute PageRank

• Simulate a random surfer?
  – On 20 billion pages and 400 billion hyperlinks...

• It can be done
  – You’ll learn how

• What about:
  – Personalized PageRank? Link spam?
What do you need to take this class?

• EECS 311
  – basic understanding of algorithms and data structures

• Helpful:
  – Basics of probability theory, machine learning

• Willingness to participate
Outline

• Introductions
• Course goals and logistics
• Why take this class?
• **What is Web Search?**
  – What will it be in five years? Ten years?
Web Search today...

- Performs an **easy** task
- Extremely **quickly**
- At massive **scale**
- Relatively **well**
Easy?

• **Belief**: Web Search engines have to understand my query and find a needle in a haystack of 20 billion documents!

• **Reality**: *Most* search queries are
  – **short** (avg. ~2.5 words [2005])
  – satisfied by pages from a **small subset** of the Web
    • Millions rather than billions [Mei et al., 2008]

=> More like finding a *pencil* in a haystack!
Extremely Quickly

• Results returned in < 1 sec
  – For any query
  – For any engine
• It was not always thus!
  – 3-4 seconds in early days ([Chu & Rosenthal, 1996; Garratt et al. 2001])
• How? Inverted indices, enormous data centers, clever algorithms...
Users are quick too...

\[ P(\text{Click at delay} < t) \]

[Downey, Dumais, & Horvitz 2007]
At Massive Scale

<table>
<thead>
<tr>
<th>Search Engine</th>
<th>February 2008</th>
<th>Searches per Day February 2008</th>
<th>February 2009</th>
<th>Searches per Day February 2009</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total</td>
<td>9882</td>
<td>341</td>
<td>13104</td>
<td>468</td>
</tr>
<tr>
<td>Google</td>
<td>5855</td>
<td>202</td>
<td>8293</td>
<td>296</td>
</tr>
<tr>
<td>Yahoo!</td>
<td>2136</td>
<td>74</td>
<td>2696</td>
<td>96</td>
</tr>
<tr>
<td>Microsoft</td>
<td>953</td>
<td>33</td>
<td>1073</td>
<td>38</td>
</tr>
<tr>
<td>Ask</td>
<td>488</td>
<td>17</td>
<td>536</td>
<td>19</td>
</tr>
<tr>
<td>AOL</td>
<td>450</td>
<td>16</td>
<td>507</td>
<td>18</td>
</tr>
</tbody>
</table>

(in millions)

Jan 2011: **18.5 Billion**  
Jan 2012: **20 Billion**

How? Inverted indices, enormous data centers, clever algorithms...

http://blog.searchenginewatch.com/comscoresearchsharefeb2009_0309.jpg
Relatively Well
Web Search Engines today...

• For the most part:
  – Perform an easy task
  – Extremely quickly
  – At massive scale
  – Relatively well

• The future of Web search is in more difficult tasks
Trend toward more difficult queries

- Search queries are getting longer (sort of)

<table>
<thead>
<tr>
<th>Subject</th>
<th>Jan-08</th>
<th>Dec-08</th>
<th>Jan-09</th>
<th>Year-over-year percent change</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 word</td>
<td>20.96%</td>
<td>20.70%</td>
<td>20.29%</td>
<td>-3%</td>
</tr>
<tr>
<td>2 words</td>
<td>24.91%</td>
<td>24.13%</td>
<td>23.65%</td>
<td>-5%</td>
</tr>
<tr>
<td>3 words</td>
<td>22.03%</td>
<td>21.94%</td>
<td>21.92%</td>
<td>0%</td>
</tr>
<tr>
<td>4 words</td>
<td>14.54%</td>
<td>14.67%</td>
<td>14.89%</td>
<td>2%</td>
</tr>
<tr>
<td>5 words</td>
<td>8.20%</td>
<td>8.37%</td>
<td>8.68%</td>
<td>6%</td>
</tr>
<tr>
<td>6 words</td>
<td>4.32%</td>
<td>4.47%</td>
<td>4.65%</td>
<td>8%</td>
</tr>
<tr>
<td>7 words</td>
<td>2.23%</td>
<td>2.40%</td>
<td>2.49%</td>
<td>12%</td>
</tr>
<tr>
<td>8+ words</td>
<td>2.81%</td>
<td>3.31%</td>
<td>3.43%</td>
<td>22%</td>
</tr>
</tbody>
</table>

Note: Data is based on four-week rolling periods (ending Jan. 31, 2009; Dec. 27, 2008; and Jan. 26, 2008) from the Hitwise sample of 10 million U.S. Internet users.

Source: Hitwise, an Experian company
Trend toward more difficult queries

- Search queries are getting longer? Sort of

<table>
<thead>
<tr>
<th>Subject</th>
<th>Jan-08</th>
<th>Dec-08</th>
<th>Jan-09</th>
<th>January 2011</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 word</td>
<td>20.96%</td>
<td>20.70%</td>
<td>20.29%</td>
<td>23.39%</td>
</tr>
<tr>
<td>2 words</td>
<td>24.91%</td>
<td>24.13%</td>
<td>23.65%</td>
<td>23.59%</td>
</tr>
<tr>
<td>3 words</td>
<td>22.03%</td>
<td>21.94%</td>
<td>21.92%</td>
<td>19.99%</td>
</tr>
<tr>
<td>4 words</td>
<td>14.54%</td>
<td>14.67%</td>
<td>14.89%</td>
<td>13.71%</td>
</tr>
<tr>
<td>5 words</td>
<td>8.20%</td>
<td>8.37%</td>
<td>8.68%</td>
<td>8.28%</td>
</tr>
<tr>
<td>6 words</td>
<td>4.32%</td>
<td>4.47%</td>
<td>4.65%</td>
<td>4.62%</td>
</tr>
<tr>
<td>7 words</td>
<td>2.23%</td>
<td>2.40%</td>
<td>2.49%</td>
<td>2.58%</td>
</tr>
<tr>
<td>8+ words</td>
<td>2.81%</td>
<td>3.31%</td>
<td>3.43%</td>
<td>3.85%</td>
</tr>
</tbody>
</table>

Note: Data is based on four-week rolling periods (ending Dec. 27, 2008; and Jan. 26, 2008) from the Hitwise sample of 10 million U.S. Internet users.

Source: Hitwise, an Experian company
Things you can’t do with search today

- Query by *description* rather than *content*
  - Humorous anecdotes about frogs

- Extracting and synthesizing over multiple pages
  - Nanotechnology companies hiring on the West Coast
  - Substances the FDA has banned

- Organizing bodies of documents
  - Show me the most compelling cases for or against QE3
  - Who says drinking from aluminum soda cans increases my Alzheimer's risk? Should I believe them?
2009 Image Search for famous pink building
Bing and Decide!

- Google “owns” algorithmic document search
## Competitive Landscape

<table>
<thead>
<tr>
<th>Core Search Entity</th>
<th>1/10: Search Share (%)</th>
<th>1/11-2-11: Explicit Core Search Share (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Jan-10</td>
<td>Jan-11</td>
</tr>
<tr>
<td>Google Sites</td>
<td>65.4%</td>
<td>65.6%</td>
</tr>
<tr>
<td>Yahoo! Sites</td>
<td>17.0%</td>
<td>16.1%</td>
</tr>
<tr>
<td>Microsoft Sites</td>
<td>11.3%</td>
<td>13.1%</td>
</tr>
<tr>
<td>Ask Network</td>
<td>3.8%</td>
<td>3.4%</td>
</tr>
<tr>
<td>AOL, Inc.</td>
<td>2.5%</td>
<td>1.7%</td>
</tr>
</tbody>
</table>

Source: ComScore, Inc.
Bing and Decide!

• Google “owns” algorithmic document search

• Bing’s strategy:
  – Searchers intend to complete **tasks**
  – Today: search engines are the **first** step
  – Goal: Have engine do the first **$k$** steps
Announcements!

• No class in this room on **Tuesday Oct 2**
  – Go to Ford ITW at 4PM to see Bryan Pardo’s talk
    • “required” for this course
• Guest lecture in class on **Thursday Oct 4**
  – Be alert for room change
• If you want to register: join the waitlist
• Sign-up for papers this week