Welcome to the class of Web Information Retrieval!
Tee Time Topic

Augmented Reality and Google Glass

By Ali Abbasi
Challenges in Web Search Engines

Min ZHANG
z-m@tsinghua.edu.cn
April 13, 2012
Challenges in Web Search Engines

- IJCAI’03 invite talk
  - Monika R. Henzinger, Google Inc. et al.
- SIGIR’05 keynote speech
  - Amit Singhal, Google Inc
- AIRS’08 keynote speech
  - Kenneth Church, Microsoft.
- WSDM’09 Keynote speech
  - Jeff Dean, Google.
- Surveys and discussions with SE companies
  - Google, Bing, Baidu, Sogou, etc.

(partial of content and several images refer to above talks / papers)
### Challenges in Web Search Engine

- **Scale, quality, quality evaluation**
- Spam
- Web conventions
- Multi sources fusion
- Evaluation
- UI
Challenges (1.1) – Scale

- 2006, 400,000,000 queries
- July 2008, 7.5 Billion queries*
- Nov. 2004, docs: 8,000,000,000
- July 25, 2008, 1 trillion docs

* Searchengine watch
http://searchenginewatch.com/3630718
Challenges (1.1) – Scale

- Peak of google.stanford.edu ~ 1997
Research Project, circa 1997

Challenges in Web IR
Challenges (1.1) – Scale

- “Corkboards” Google servers 1999
Serving System, circa 1999

Challenges in Web IR
Challenges (1.1) – Scale

- Google Datacenter 2000
Challenges (1.1) – Scale

- Google new datacenter 2001
Challenges (1.1) – Scale

- Google new datacenter 2001 (3 days later)
Early 2001, in-memory index

Holding the complete search index in memory: resulting in the use of 1000 machines to handle a single query compared to just 12 previously

Search latency:
1000ms → 200ms
Challenges (1.1) – Scale

- Current large scale computing

wave cooling system, California
Server design 2004

Diagram of server design with Cache servers, Requests, Root, Parent Servers, Repository Manager, Leaf Servers, and GFS.
2007 universal search

Challenges in Web IR
Late 2009: new 'Caffeine' engine

11, Aug, 2009 “Google unveils new "Caffeine" search engine.”

- Just after the partnership between Yahoo! and Bing
- To make the search process much faster, more intuitive and overall produce results with greater accuracy for the user.
- The new architecture is said to include size, indexing, speed, accuracy, crawling and ranking changes.

Some conclusions:

- Returns results at over double the speed to the current Google.
- Places more reliance on keyword strings rather than exact phrases.
- Gives greater weight to authority domains and social media, e.g. blogs
- Gives increased weight on domain names
- Have more accurate real time results in an effort to match Facebook and Twitter’s real time search engines.
Challenges (1.1) – Scale

Beyond system architecture:

- IR methods need significant rethinking
  - **idf**
    - How meaningful is idf in a heterogeneous collection of this size?
    - All words have very high idf
  - **Stemming**
    - Often used as a recall tool
    - Do you really need recall when you have 1 trillion docs?
  - **Query expansion**
    - Is it at all needed? Does it work? What about query drift?
    - How fast can you do it? What is the effect of slowdown on user?
Challenges in Web IR

### Challenges (1.2) – Scale, and quality

<table>
<thead>
<tr>
<th>Search Engine</th>
<th>Reported index Size</th>
<th>Page Depth</th>
</tr>
</thead>
<tbody>
<tr>
<td>Google</td>
<td>8.1 billion (Dec. 2004)</td>
<td>101K</td>
</tr>
<tr>
<td>MSN</td>
<td>5.0 billion</td>
<td>150K</td>
</tr>
<tr>
<td>Yahoo!</td>
<td>19.2 billion (Aug. 2005)</td>
<td>500K</td>
</tr>
<tr>
<td>Ask Jeeves</td>
<td>2.5 billion</td>
<td>101K+</td>
</tr>
<tr>
<td>All the Web</td>
<td>152 billion</td>
<td>605K</td>
</tr>
<tr>
<td>All the Surface Web</td>
<td>10 billion</td>
<td>8K</td>
</tr>
</tbody>
</table>

**Battle for the index**

**“Absolute numbers are no longer useful”** -- by Google, Sep 27, 2005

- How big is the Web? – 5 billion is enough
  -- by Kenneth Church (Microsoft Research), Jan, 2008

From Danny Sullivan, SearchEngineWatch web site

2002.12
Challenges (1.2) — quality

- Misleads its human readers as well
  - Wrong info
    - e.g. 1st president of USA is Thomas Jefferson,
  - Misleading medical information
  - Once correct, but out of date
    - e.g. election result, #hurt in one accident

- Perhaps a good start point
  - Link analysis
    - PageRank, HITS, BrowseRank, …
Challenges (1.2) — quality

- An interesting point —
  Evaluating the quality of *anchor text*
  - High quality pages → low quality anchor text?
  - Judge the quality of anchor text *independently*?
  - Anchor text: editorial or purely descriptive?
  - To identify *multiple themes* in documents?

- A promising area of research —
  *established quality judgments*
  *(link + text + user behavior) based analysis*
Challenges (1.2) – quality evaluation

- Users are reluctant to give direct feedback
- Using log data
  - Still incomplete
  - User’s click = the page is useful
- Collect the click-through data
  - Under some weak assumptions
    - Click-through of ranking A > Click-through of ranking B
      If and only if A shows more relevant info than B
- A necessary and important rising research field
  - User behavior reliability analysis
Current progress: Finding high quality pages using query-independent features

<table>
<thead>
<tr>
<th>Data set</th>
<th>English (.gov)</th>
<th>Chinese (.com)</th>
</tr>
</thead>
<tbody>
<tr>
<td># pages</td>
<td>1,247,753</td>
<td>37,205,218</td>
</tr>
<tr>
<td>Total size</td>
<td>18.1G</td>
<td>558.0 G</td>
</tr>
<tr>
<td>#Queries / #target pages</td>
<td>100 / 2631</td>
<td>650 / 48930</td>
</tr>
<tr>
<td>Found high-quality pages</td>
<td>52.00%</td>
<td>4.96%</td>
</tr>
<tr>
<td>Target pages recall (training set)</td>
<td>95.53%</td>
<td>92.73%</td>
</tr>
<tr>
<td>Target pages recall (test set)</td>
<td>93.57%</td>
<td>92.37%</td>
</tr>
</tbody>
</table>
Challenges in Web Search Engine

- Scale, quality, quality evaluation
- Spam
- Web conventions
- Multi sources fusion
- Evaluation
- UI

To be continued...