Search Quality

Jan Pedersen
10 September 2007
Outline

- The Search Landscape
- A Framework for Quality
  - RCFP
- Search Engine Architecture
- Detailed Issues
Three major “Mainframes”
  – Google, Yahoo, and MSN

>800M searches daily
  – 60% international
  – $10^6$ machines

$20B$ in Paid Search Revenues

Large indices
  – Billions of documents
  – Petabytes of data

Key Drivers:

Scale, Quality, Distribution
Search Engine Architecture

WWW

Crawling

Stream Computation

WebMap

Meta data

Indexing

Snapshot

Load Replication

Query Serving

CPU Bottleneck

Bandwidth Bottleneck

Index
**Query Serving Architecture**

- **Rectangular Array**
  - Each row is a replicate
  - Each column is an index segment

- Results are merged across segments
  - Each node evaluates the query against its segment.

- Latency is determined by the performance of a single node

---

<table>
<thead>
<tr>
<th>Sunnyvale</th>
<th>NYC</th>
</tr>
</thead>
<tbody>
<tr>
<td>L3</td>
<td>L3</td>
</tr>
</tbody>
</table>

```
<table>
<thead>
<tr>
<th></th>
<th>FE_1</th>
<th>FE_2</th>
<th>...</th>
<th>FE_32</th>
</tr>
</thead>
<tbody>
<tr>
<td>QI_1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>QI_2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>FI_1</td>
<td>FI_2</td>
<td>...</td>
<td>FI_48</td>
</tr>
</tbody>
</table>

```

```
<table>
<thead>
<tr>
<th></th>
<th>WI_1,1</th>
<th>WI_1,2</th>
<th>WI_1,3</th>
<th>...</th>
<th>WI_1,48</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>WI_2,1</td>
<td>WI_2,2</td>
<td>WI_2,3</td>
<td>...</td>
<td>WI_2,48</td>
</tr>
<tr>
<td></td>
<td>WI_3,1</td>
<td>WI_3,2</td>
<td>WI_3,3</td>
<td>...</td>
<td>WI_3,48</td>
</tr>
<tr>
<td></td>
<td>WI_4,1</td>
<td>WI_4,2</td>
<td>WI_4,3</td>
<td>...</td>
<td>WI_4,48</td>
</tr>
</tbody>
</table>

```

---
What’s the Goal?

- User Satisfaction
  - Understand user intent
    - Problems: Ambiguity and Context
  - Generate relevant matches
    - Problems: Scale and accuracy
  - Present useful information
    - Problems: Ranking and Presentation
Evaluation

- Graded Relevance score
- Editorial Assessment
- Session/Task fulfillment?
- Behavioral measures?
Average highest rank clicked perceptibly increased with the release of a new rank function.
Quality Dimensions

- Ranking
  - Ability to rank hits by relevance
- Comprehensiveness
  - Index size and composition
- Freshness
  - Recency of indexed data
- Presentation
  - Titles and Abstracts
Comprehensiveness

- **Problem:**
  - Make accessible all useful Web pages

- **Issues:**
  - Web has an infinite number of pages
  - Finite resources available
    - Bandwidth
    - Disk capacity

- **Selection Problem**
  - Which pages to visit
    - Crawl Policy
  - Which pages to index
    - Index Selection Policy
Moore’s Law and Index Size

- ~150M in 1998
- ~5B in 2005
  - 33x increase
  - Moore would predict 25x
- What about 2010?
  - 40B?

Source: Search Engine Watch

- 1994 Yahoo (directory) and Lycos (index) go public
- 1995 Infoseek and Excite go public
- 1997 Alta Vista launches 100M index
- 1998 Inktomi and Google launched
- 1999 All The Web launched
- 2003 Yahoo purchases Inktomi and Overture
- 2004 Google goes public
- 2005 Msft launches MS Live
Freshness

- Problem:
  - Ensure that what is indexed correctly reflects current state of the web
- Impossible to achieve exactly
  - Revisit vs Discovery
- Divide and Conquer
  - A few pages change continually
  - Most pages are relatively static
Changing documents in daily crawl for 32-day period
Freshness

Source:
Search Engine Showdown
Ranking

- **Problem:**
  - Given a well-formed query, place the most relevant pages in the first few positions

- **Issues:**
  - Scale: Many candidate matches
    - Response in < 100 msecs
  - Evaluation:
    - Editorial
    - User Behavior
 Ranking Framework

- **Regression problem**
  - Estimate editorial relevance given ranking features

- **Query Dependent features**
  - Term overlap between query and
    - Meta-data
    - Content

- **Query Independent Features**
  - Quality (e.g. Page Rank)
  - Spamminess
Goal: Automatically construct a ranking function
  – Input:
    • Large number training examples
    • Features that predict relevance
    • Relevance metrics
  – Output:
    • Ranking function

Enables rapid experimental cycle
  – Scientific investigation of
    • Modifications to existing features
    • New feature
Ranking Features

- A0 - A4: anchor text score per term
- W0 - W4: term weights
- L0 - L4: first occurrence location (encodes hostname and title match)
- SP: spam index: logistic regression of 85 spam filter variables (against relevance scores)
- F0 - F4: term occurrence frequency within document
- DCLN: document length (tokens)
- ER: Eigenrank
- HB: Extra-host unique inlink count
- ERHBER*HB
- A0W0 etc.: A0*W0
- QA: Site factor – logistic regression of 5 site link and url count ratios
- SPN: Proximity
- FF: family friendly rating
- UD: url depth
Implements (Tree 0)

\[ A_{0w0} < 22.3 \]

\[ L_0 < 18+1 \]

\[ L_1 < 18+1 \]
\[ R = -0.0545 \]

\[ W_0 < 856 \]
\[ R = -0.0039 \]
\[ F_2 < 1 + 1 \]
\[ R = -0.1604 \]

\[ L_1 < 509+1 \]
\[ R = -0.0199 \]
\[ F_0 < 2 + 1 \]
\[ R = -0.0199 \]
\[ R = -0.1185 \]

\[ N \]
\[ R = 0.0015 \]
• Spelling Correction
• Also Try
• Short cuts
• Titles and Abstracts
Eye Tracking Studies

- Golden Triangle
  - Top left corner
- Quick scan
  - For candidate
- Longer scan
  - For relevance
Comparison to State-of-the-art

Comparison to State-of-the-art

**Niangua Darter**

... Measures taken to stabilize and improve Niangua darter habitat will also benefit...

from fertilizers and pesticides threaten Niangua darter habitat...

www.conservator.state.mo.us/natis/endangered/endangered/darter/...

**Arkansas Darter**

... and general development resulted in major losses of Arkansas darter habitat.

Since the late 19th century, the Arkansas darter's habitat has been reduced...

www.conservator.state.mo.us/natis/endangered/endangered/arkdarter/...

**Relict Darter** Ethostoma cheniensis U.S. Fish & Wildlife Service

... of the Bayou du Chien that has significantly altered the darter's habitat.

This massive alteration of the relict darter's habitat reduced both...

endangered.fws.gov/e/esa38.html - 10k - Cached - Similar pages

**Slackwater Darter** Ethostoma boschungi U.S. Fish & Wildlife Service

... of the slackwater darter varies with the temperature of the breeding habitat and...

slackwater darter habitat in the Cypress Creek drainage, Tennessee,...

endangered.fws.gov/e/esa1a.html - 9k - Cached - Similar pages

More results from endangered.fws.gov

**Characteristics of Channel Darter Habitat in the Winoski River**

File Format: Microsoft Powerpoint 97 - View as HTML

Characteristics of Channel Darter Habitat in the Winoski River, Vermont. Douglas E. Facey and Shannon M. O'Brien. Department of Biology...

academics.smccd.edu/facdev/Scholarship/AFS%202003%20poster.ppt - Similar pages

**Natural Heritage Program - The Maryland Darter**

... We will probably never know because the Maryland Darter's habitat is the point...

This change in the Maryland Darter's habitat would have been a major...

www.dnr.state.md.us/wildlife/mddarter.asp - 10k - Apr 7, 2005 - Cached - Similar pages

**UA AFS's Position on Ark. Darter Habitat Preservation**
Conclusions

- Search is a hard problem
  - Solutions are approximate
  - Measurement is difficult

- Search quality can be decomposed into separate but related problems
  - Ranking
  - Comprehensiveness
  - Freshness
  - Presentation