Web Search

Interfaces
Web Search Interface

- Web search engines of course need a web-based interface.
- Search page must accept a query string and submit it within an HTML `<form>`.
- Program on the server must process requests and generate HTML text for the top ranked documents with pointers to the original and/or cached web pages.
- Server program must also allow for requests for more relevant documents for a previous query.
Submit Forms

- HTML supports various types of program input in forms, including:
  - Text boxes
  - Menus
  - Check boxes
  - Radio buttons

- When user submits a form, string values for various parameters are sent to the server program for processing.

- Server program uses these values to compute an appropriate HTML response page.
Simple Search Submit Form

<p> <b> Enter your query: </b> </p>
<input type="text" name="query" size=40>

<p> <b>Search Database: </b> </p>
<select name="directory">
<option selected value="/u/mooney/ir-code/corpora/cs-faculty/"> UT CS Faculty </option>
<option value="/u/mooney/ir-code/corpora/yahoo-science/"> Yahoo Science </option>
</select>

<p> <b>Use Relevance Feedback: </b> </p>
<input type="checkbox" name="feedback" value="1">
<br> <br>
<input type="submit" value="Submit Query">
<input type="reset" value="Reset Form">
</form>
Java Servlet

- Java’s approach to processing web form requests.
- Program runs on Web server and builds pages on the fly.
- Servlet code supporting sample interface is in
  - /u/mooney/ir-code/irs/
Simple Search Servlet

- Based on directory parameter, creates or selects existing InvertedIndex for the appropriate corpus.
- Processes the query with VSR to get ranked results.
- Writes out HTML ordered list of 10 results starting at the rank of the start parameter.
- Each item includes:
  - Link to the original URL saved by the spider in the top of the document in BASE tag.
  - Name link with page <TITLE> extracted from file.
  - Additional link to local cached file.
- If all retrievals not already shown, creates a submit form for “More Results” starting from the next ranked item.
Simple Search Interface Refinements

• For “More results” requests, stores current ranked list with the user session and displays next set in the list.
• Integrates relevance feedback interaction with “radio buttons” for “NEUTRAL,” “GOOD,” and “BAD” in HTML form.
Other Search Interface Refinements

- Highlight search terms in the displayed document.
  - Provided in cached file on Google.
- Allow for “advanced” search:
  - Phrasal search (“..”)
  - Mandatory terms (+)
  - Negated term (-)
  - Language preference
  - Reverse link
  - Date preference
Clustering Results

- Group search results into coherent “clusters”:
  - “microwave dish”
    - One group of on food recipes or cookware.
    - Another group on satellite TV reception.
  - “Austin bats”
    - One group on the local flying mammals.
    - One group on the local hockey team.
- **Northern Light** used to group results into “folders” based on a pre-established categorization of pages (like DMOZ categories).
- Alternative is to dynamically cluster search results into groups of similar documents.
User Query Length

• Users tend to enter short queries.
  – Study in 1998 gave average length of 2.35 words.
• Evidence that queries are getting longer.

<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
Speech Queries are Longer

Microsoft internal data April 2015