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

```
<form action="http://prospero.cs.utexas.edu:8082/servlet/irs.Search" method="POST">
<p> <b> Enter your query: </b>
    <input type="text" name="query" size=40>
<p> <b>Search Database: </b>
   <select name="directory">
   <option selected value="/u/mooney/ir-code/corpora/cs-faculty/"> UT CS Faculty
   <option value="/u/mooney/ir-code/corpora/yahoo-science/"> Yahoo Science
  </select>
<p> <b>Use Relevance Feedback: </b>
<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
- Machine translation of pages.

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.

Percentage of U.S. clicks by number of keywords				
Subject	**Jan-08**	**Dec-08**	**Jan-09**	**Year-over-year percent change**
1 word	20.96%	20.70%	20.29%	**-3%**
2 words	24.91%	24.13%	23.65%	**-5%**
3 words	22.03%	21.94%	21.92%	**0%**
4 words	14.54%	14.67%	14.89%	**2%**
5 words	8.20%	8.37%	8.68%	**6%**
6 words	4.32%	4.47%	4.65%	**8%**
7 words	2.23%	2.40%	2.49%	**12%**
8+ words	2.81%	3.31%	3.43%	**22%**
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