Web Search and linguistics

- Web search is important
  - 85% use search engines
  - 1/3 of web sessions involve search engines
- Web search is difficult
  - Large volumes of data
  - Dead links, dynamic pages, relocations, frames, etc.
  - Variable document quality
  - Multitude of languages

- Too many documents available
- How to select the relevant documents
- How to rank the selected documents

Using linguistic techniques to analyze document content
Outline of Presentation

- Why is Web Search so difficult?
- Why is linguistics needed?
- Internet Search Architecture and AllTheWeb
- Linguistic Techniques on AllTheWeb
- Discussion

Why is Web Search so Difficult?

- Volume of data:
  - Document explosion (table from 2005)
  - Document dynamics
  - Distributed over many computers and platforms

<table>
<thead>
<tr>
<th>Search Engine</th>
<th>Reported Size</th>
<th>Page Depth</th>
</tr>
</thead>
<tbody>
<tr>
<td>Google</td>
<td>8.1 billion</td>
<td>101K</td>
</tr>
<tr>
<td>MSN</td>
<td>5.0 billion</td>
<td>150K</td>
</tr>
<tr>
<td>Yahoo</td>
<td>4.2 billion (estimate)</td>
<td>500K</td>
</tr>
<tr>
<td>Ask Jeeves</td>
<td>2.5 billion</td>
<td>101K+</td>
</tr>
</tbody>
</table>

- Multitude of languages:
  - Multi-lingual web
  - 40-50 languages used on the web
  - Many text encoding standards
Why is Web Search so Difficult?

- **Document Quality:**
  - Misspellings
  - Spam and offensive content
  - Little text
  - All topics

- **User Behavior:**
  - Misspellings
  - Query length: 2.4 terms
  - Query session: 8 queries
  - Half of the documents viewed are among top three documents on result page

<table>
<thead>
<tr>
<th>Query</th>
<th>No. of documents</th>
</tr>
</thead>
<tbody>
<tr>
<td>événements</td>
<td>76,000</td>
</tr>
<tr>
<td>événements</td>
<td>420,000</td>
</tr>
<tr>
<td>événements</td>
<td>35,000</td>
</tr>
<tr>
<td>événements</td>
<td>95,000</td>
</tr>
<tr>
<td>événements</td>
<td>22,000</td>
</tr>
<tr>
<td>événements</td>
<td>9,000</td>
</tr>
</tbody>
</table>

Why is Linguistics Needed?

- **Ultimate goal**
  - Understand document content
  - Understand user information need

- **Immediate goal**
  - Handle syntactic variation
  - Handle morphological variation

- **Effects**
  - More relevant documents retrieved
  - Most relevant document on top of list
  - Result list more understandable to user
Web Search Architecture

- AllTheWeb search architecture (2002)

Crawling cycle: 9-11 days
Online news sources continuously crawled

800 mill. pages (Oct 2003: 3.1 billion)
2000 online news sources
115 mill. pictures
2 mill. MP3 songs

(150,000 unique visitors every week)
30 mill. queries per day

Web Search Architecture

Linguistic Techniques

- Three categories of linguistic/text mining techniques

Categorizing techniques

<non> Search options
- All selected
- Category-based selection
- Documents Categories of documents

Transformational techniques

Query Transformed query
Keyword-based search
Content-based search
Relevant documents Transformed documents

Presentational techniques

- Title-based access
- Content-based access

List of documents Presentation of document list

Improved transparency

- Speed
- Flexibility (modifications and languages)

- Increased semantics
- Reduced search space
Linguistic techniques: 1

- Language Identification
  - Allows the user to select the language of the retrieved documents
  - Method:
    
    I. Languages with clear word boundaries: Identify language by dictionary lookup
    II. Languages with no clear word boundaries: Identify language by checking document against frequency list of bigrams
    III. Additional strategies:
      - HTML structural information and meta information
      - Domain names (e.g., .NO)

- Status (Jan 2002):
  - 52 languages recognized
  - 95-96% of web documents successfully tagged for language

Linguistic Techniques: 2

- Offensive Content Reduction
  - Allows the user to filter out offensive documents (e.g., pornography)
  - Method:
    
    I. Maintain offensive dictionary of weighted words and phrases
    II. Traverse first part of document and calculate score for offensive material based on dictionary lookup
    III. Tag offensive documents in index

- Status:
  - English, German, Italian, Spanish, and French
  - Problematic for documents with little text
Linguistic Techniques: 3

• **Text Categorization**

  - Allows the user to restrict the search to certain categories
  - Each category is defined by a dictionary with characteristic words and phrases
  - Example: SCIRUS library of scientific information (www.scirus.com)
  - Categorization dimensions:
    - content
    - type

  - Categorization difficult for the Web

---

**Method:**

I. Maintain dictionary of words and phrases with associated categories
II. Content categorization: Calculate score for each category by looking up words and phrases in the dictionary
III. Extract meta information from structural information in document
IV. Type categorization: Deduce document type using rules that analyze the document's meta information

- Categories: computer science, economics, business and management, etc.
- Types: homepage, abstract, article, etc.
Linguistic Techniques: 4

• Lemmatization

  – Map inflections of words onto one canonical representation (lemma)

    – Examples:
      writes -> write
      written -> write
      was -> be
      cars -> car
      better -> good
      best -> good

  – Retrieve documents also if inflections do not match; semantic retrieval

    Several ways of implementing lemmatization

    
    | Query          | Document       |
    |----------------|----------------|
    | Actual text    | Lemmatized text|
    | good cars      | good car       |
    | The best car...| The good car...|

• Lemmatization cont.

  – Method:

    | Strategy                      | Query     | Document   |
    |--------------------------------|-----------|------------|
    | Expand with all forms         | car       | cars       |
    | Expand with base form         | cars      | car        |
    | Replace with base form        | car       | car        |

  – Status:

    – Need large full-form dictionaries during indexing (e.g. around 1 million full-forms in German dictionary)
    – Increase of Russian/Polish index: 600-800%
    – Increase of German index: 5%
Linguistic Techniques: 5

- **Phrasing**
  - Some queries should be interpreted as phrases to increase precision
    - mutual information
      - 1,000,000 documents
    - "mutual information"
      - 6,300 documents
  - Phrases recognized in documents, queries and indexes
    - Method:
      1. Maintain phrase list during querying (English, German, French)
      2. Identify and quote longest left-most phrases in query from list of phrases
      3. Add bigrams of all consecutive query terms that are not part of a phrase as optional phrases
  - Example:
    - New York art museum
      - ALL
      - +"New York" +art +museum "art museum"

Linguistic Techniques: 6

- **Anti-Phrasing**
  - Some queries contain phrases that disturb the search
    - where can I find The Economist
      - Homepage of The Economist not found
    - The Economist
      - Homepage of The Economist found
  - Anti-phrasing means to remove irrelevant phrases in the beginning of queries
    - Method:
      1. Maintain anti-phrase list during querying (English, German, French)
      2. Phrase query
      3. Identify and remove longest left-most anti-phrase starting at position 1 in query
  - Example:
    - where do I find New York
      - Phrasing
      - where do I find "New York"
      - Anti-phrasing
      - "New York"
Linguistic Techniques: 7

• Document Clustering

  – Present list of retrieved documents in the form of a hierarchical tree
  – Clusters computed on the fly based on X highest ranked documents
  – Clusters sum up possible topics of query
  – Method:

    I. Retrieve X highest ranked documents for the query posted
    II. For each document:
      a. Extract words and phrases and assign numeric measure of importance to each of them
      b. Construct document vector (or use precomputed vector)
      c. For clusters already known: Try to map document to existing clusters
      d. If mapping to existing cluster fail: Construct new cluster on the fly and assign document to new cluster

  – Clustering is language-independent (except for stopword lists)

Discussion

• Speed and Flexibility
  – Finite-state automata
  – Dictionary-driven applications

• Relevance Effects
  – Difficult to measure

• Future Challenges
  – Linguistic sophistication vs. speed/space requirements
  – Still far from real semantic information retrieval
  – Semantic Web