Extending/Enhancing Model

Revisit: Modeling

• Define each of
  – Information object
  – Query
  – Satisfying of query by objects
  – Useful presentation:
    • focus on ranking

Models have seen

<table>
<thead>
<tr>
<th>Model</th>
<th>Document</th>
<th>Query</th>
<th>Satisfy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Boolean</td>
<td>set of terms</td>
<td>boolean expression over terms</td>
<td>evaluate boolean expression</td>
</tr>
<tr>
<td>Vector</td>
<td>dictionary of t terms</td>
<td>t-dimensional vector</td>
<td>t-dimensional vector</td>
</tr>
<tr>
<td>Extend</td>
<td>more information ?</td>
<td>more information ?</td>
<td>?</td>
</tr>
</tbody>
</table>

Start to enhance model

Properties of terms within documents

– Frequency of term in doc
– Where in doc?
– Relative position of multiple query terms
– Special use? (e.g. in title, font, … )
– Occurs in anchor text of another doc, pointing to this doc.

Start to enhance model

Properties of terms within documents

Vector model gave us
– Frequency of term in doc

Property of each occurrence of term in doc.
– Where in doc?
  – Relative position of multiple query terms
  – Special use? (e.g. in title, font, … )

Found when evaluate another document
– Occurs in anchor text of another doc. pointing to this doc.

Model

• Document: bag of terms + attributes
  – Includes sequence of terms. Why?
• Query: sequence of terms
  – Can make more complicated
• Satisfying: in current search engines, documents "containing" all terms
  – AND model
    – "containing" includes anchor text of pointers to this doc from other docs
    – Sometimes relax if no documents contain all
• Ranking: wide open function
  – info beyond documents and query?
**Data Structure for Collection**

- for each document, keep list of:
  - **terms** appearing
  - **positions** at which each term occurs
  - **attributes** for each occurrence of term
- keep summary information for documents

**Data Structure for Collection: Invert**

- for each term, keep list of:
  - **documents** in which it appears
  - **positions** at which it occurs in each doc.
  - **attributes** for each occurrence
- keep summary information for documents
- keep summary information for terms

**Inverted Index for Collection**

- for each term, keep **POSTINGS LIST** of:
  - each **document** in which it appears
  - each **position** at which it occurs in doc.
    - **attributes** for each occurrence
- Core structure used by query evaluation and document ranking algorithms

**Index structure**

```
term_1: (doc ID (position, attributes))
  (position, attributes),
  (position, attributes)
  ...
(doc ID)
  (position, attributes),
  ...
  (position, attributes)
...
term_2: (doc ID (position, attributes))
  (position, attributes),
  ...
  (position, attributes)
...
```