Lecture 7:
Term Relationships & Grouping
Problems with Single-Term Indexing

- Single terms are either too specific or too broad
- Single terms carry no context
- Single terms are more ambiguous
Generation of Complex Identifiers

• Manual content analysis and indexing

• Automatic

  Linguistic analysis (to generate linguistically related terms)

  Term clustering (based on term co-occurrence stats)

  Probabilistic analysis (incorporating term-dependence information)
Automatic Term Classification

- Construct term matrix from existing document collection

<table>
<thead>
<tr>
<th></th>
<th>$T_1$</th>
<th>$T_2$</th>
<th>...</th>
<th>$T_t$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$D_1$</td>
<td>$d_{1,1}$</td>
<td>$d_{1,2}$</td>
<td>...</td>
<td>$d_{1,t}$</td>
</tr>
<tr>
<td>$D_2$</td>
<td>$d_{2,1}$</td>
<td>$d_{2,2}$</td>
<td>...</td>
<td>$d_{2,t}$</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>$D_n$</td>
<td>$d_{n,1}$</td>
<td>$d_{n,2}$</td>
<td>...</td>
<td>$d_{n,t}$</td>
</tr>
</tbody>
</table>

- Similar terms tend to be used in the same documents:

  Group terms based on similarity amongst columns

- Similar documents contain related terms:

  Group docs into doc classes based on similarity between rows, then group terms with high frequency of co-occurrence within a doc class
Problems

• Co-occurring terms may not be related!

• Statistical methods may not be reliable (low precision and recall)
Linguistic Methods

• Identify syntactic classes and construct word phrases based on patterns of syntactic markers (such as noun-noun, adjective-noun)

• Problems:

  Ambiguous words and syntactic structures

  Unreliable

• Solution:

  Develop good parser/semantic analysers

  Use statistical methods to resolve ambiguity

  Accept fact that automatic analysis is not perfect
Term Phrase Formation

- Provides more specific information than single terms, e.g.:

1. Choose a phrase head (high freq term or term with negative discriminatory value)
2. Add to this other terms with low/medium frequency (can limit terms to occur in same sentence, etc)
3. Eliminate stop words

The more restrictions in step 2, the fewer phrases

- Can combine with linguistic analysis. Term phrases:

must conform to specific syntactic patterns
must occur within same sentence unit
can be augmented with domain-specific semantic analysis
conceptual graphs (semantically similar, but syntactically different)
Thesaurus Group Generation

- Thesaurus can be used to broaden scope of terms

- Can convert every term in same class to the name of the class (controlled vocabulary)

- Can also stem to reduce size of thesaurus (but must ensure that different word senses are maintained)

- Domain-specific thesauri are usually created manually
• Thesaurus Group Generation based on term co-occurrence

Given the term-document matrix:

<table>
<thead>
<tr>
<th></th>
<th>$T_1$</th>
<th>$T_2$</th>
<th>...</th>
<th>$T_t$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$D_1$</td>
<td>$d_{1,1}$</td>
<td>$d_{1,2}$</td>
<td>...</td>
<td>$d_{1,t}$</td>
</tr>
<tr>
<td>$D_2$</td>
<td>$d_{2,1}$</td>
<td>$d_{2,2}$</td>
<td>...</td>
<td>$d_{2,t}$</td>
</tr>
<tr>
<td></td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>$D_n$</td>
<td>$d_{n,1}$</td>
<td>$d_{n,2}$</td>
<td>...</td>
<td>$d_{n,t}$</td>
</tr>
</tbody>
</table>

Compute the similarity between terms $T_j$ and $T_k$:

$$sim(T_j, T_k) = \frac{\sum_{i=1}^{n} d_{i,j} d_{i,k}}{\sqrt{\sum_{j=1}^{n} d_{i,j}^2 \times \sum_{i=1}^{n} d_{i,k}^2}}$$

Single-link classification: 2 words are put into same group if sim > threshold
Complete-link: sim of each pair of words in a group > threshold
Pseudo Classification

• Given a sample collection, and a sample set of queries with relevance judgements:

    if $D$ and $Q$ are judged relevant, two terms $T_j$ in $Q$ and $T_k$ in $D$ are placed in same group

    Such assignment will increase sim between $D$ and $Q$

• Similar principle is used in relevance feedback