Contents

1 Term Relationships and Grouping

1.1 Problems with Single-Term Indexing

- Single terms are either too specific or too broad.
- Single terms carry no context.
- Single terms are more ambiguous.

1.2 Generation of Complex Identifiers

- Manual content analysis and indexing.
- Automatic:
  - Linguistic analysis to generate linguistically related terms.
  - Term clustering based on term co-occurrence statistics.
  - Probabilistic analysis incorporating term-dependence information.
    Estimation of joint occurrence probabilities for pairs and triplets of words: Too expensive yet unreliable accuracy.

1.3 Automatic Term Classification

- Construct a term-document matrix from an existing document collection:

	T_1	T_2	...	T_t
D_1	$d_{1,1}$	$d_{1,2}$	...	$d_{1,t}$
D_2	$d_{2,1}$	$d_{2,2}$	...	$d_{2,t}$
...	...	...	...	...
...	...	...	...	...
D_n	$d_{n,1}$	$d_{n,2}$	...	$d_{n,t}$

- Similar terms tend to be used in the same documents: Group terms together based on similarity among columns.

- Similar documents contain related terms: Group documents into document classes based on similarity between rows, and then group terms that cooccur frequently with a document class.

1.3.1 Problems

  - Co-occurring terms not necessarily related: relation may be local to a collection of documents.
  - Statistical methods may not be reliable (as reflected in low precision and recall).

1.4 Linguistic Methods

- Identification of syntactic classes and construct word phrases based on patterns of the syntactic marker (e.g, noun-noun, adjective-noun).
- Problems: Ambiguous words and syntactic structures, unreliable. Note syntactic structures can't be simply lookup from a dictionary (contrarily to what the textbook said).
- Solution:
  - Develop good parsers and semantic analyzer
  - Use statistical methods to resolve ambiguity
  - Accept the fact that the automatic analysis can never be perfect.

1.5 Term Phrase Formation

- Term phrases provide more specific information than single terms.
- Simple phrase-formation process:
  1. Choose a phrase head — a term with a high document frequency or negative discrimination value.
  2. Add to the phrase head other medium or low frequency terms. Restriction can apply: must occur within the same sentence, same grammatical unit, or within a certain proximity.
  3. Elimination of function words.
- Restrictions in Step 2 can lead to more or fewer term phrases.
- Combine with linguistic analysis:
  - Term phrases must conform to certain syntactic pattern (e.g., noun-noun).
  - Term phrases must occur within the same sentence units (e.g., subject phrase, object phrase, verb phrase).
  - Augmented with domain-specific semantic analysis (e.g., to detect technical terms within a field).
  - Recognition of semantically identical but structurally different phrases during indexing. Alternatively, generation of semantically equivalent phrase structures from phrases specified in the query.

1.6 Thesaurus Group Generation

- Thesaurus can be used to broaden the scope of a term.
- Convert every terms within the same class to the name of class.
- Stemming can be applied at the same time to reduce the size of the thesaurus.
- Thesaurus are typically constructed manually for a particular domain (e.g., CACM Computing Review Classification).

1.7 Thesaurus Group Generation Based on Term Co-occurrence

- Given a term-document matrix for an existing document collection:

	T_1	T_2	...	T_t
D_1	$d_{1,1}$	$d_{1,2}$	...	$d_{1,t}$
D_2	$d_{2,1}$	$d_{2,2}$	...	$d_{2,t}$
...	...	...	...	...
...	...	...	...	...
D_n	$d_{n,1}$	$d_{n,2}$	...	$d_{n,t}$

- Compute the *similarity* between terms T_j and T_j:

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

- Single-link classification: two words are put into the same group if their similarity exceeds a threshold.
- Complete-link classification: the similarity of *each pair* of words with a group must exceed a threshold.

1.8 Pseudo Classification

- Given a sample collection, a sample set of queries with relevance judgement, if D and Q are judged to be relevant, two terms T_j in Q and T_k in D are put into the same group.
- Such assignment will increase the similarity between D and Q.
- Similar principle used in relevance feedback.
- Query driven: examines a small number of documents; unreliable to generalize from a few queries.