Collocations and Information Management Applications

Gregor Erbach
Saarland University
Saarbrücken

Outline

- Information Management Applications
- Information Retrieval Techniques
- Categorization, Clustering
- Summarization
- Information Extraction
- Question Answering
- Points for discussion

Well-known Applications of Collocations

- Lexicography
- Machine Translation
- NL Generation
- NL Parsing
- Terminology Extraction
- Foreign Language Teaching
- Speech Recognition

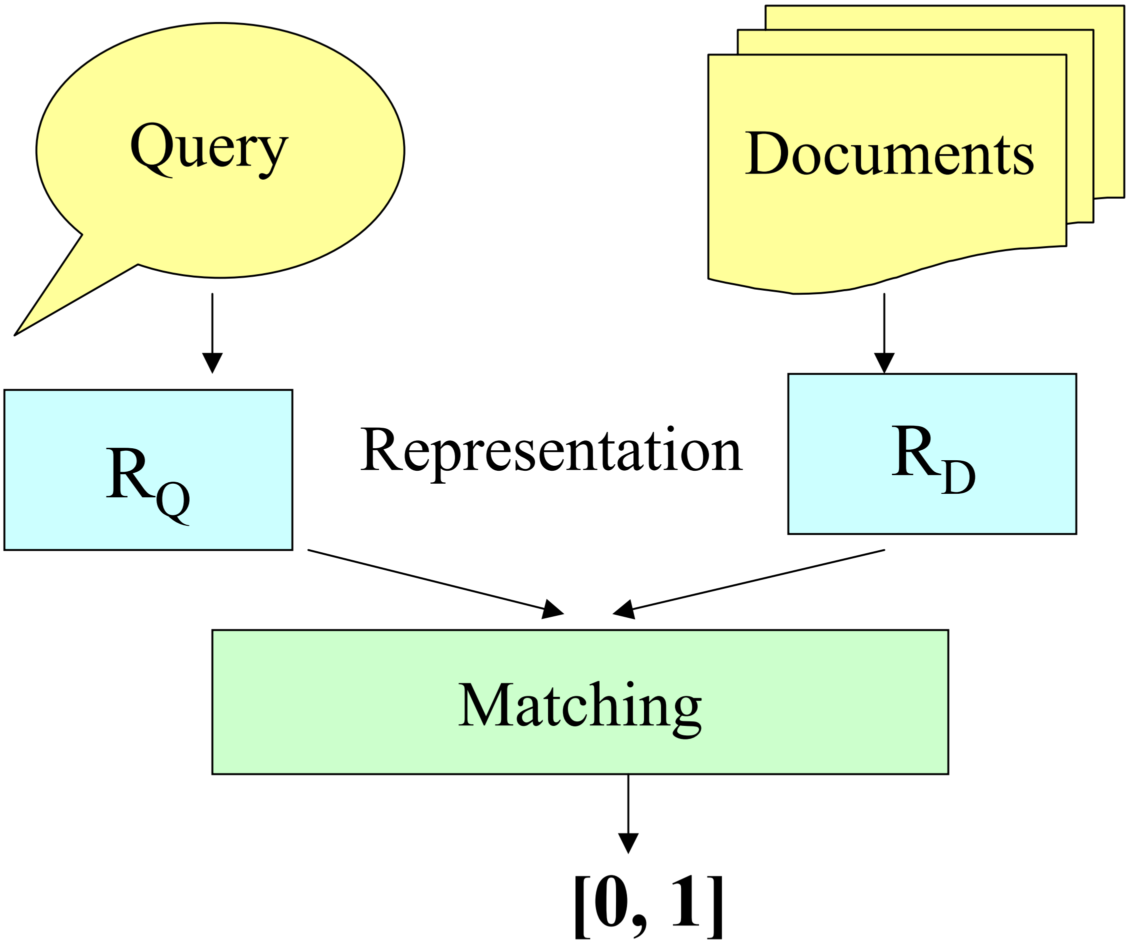
Information Management Applications

- Information Retrieval
- Text Categorisation
  (by language, topic, author, genre ...)
- Clustering
- Summarisation / Keyword Extraction
- Information Extraction
- Question Answering

Information Retrieval

- Most IR systems don't retrieve information, but documents
- Boolean retrieval: an unordered set of documents are returned as result for a query
- Ranked retrieval: an ordered list of documents is returned; relevance of documents is determined by matching with a query

IR System Model



Query Languages

- Co-occurence within document
  information AN D retrieval
- Negation
  information AND (NOT retrieval)
- Multi-word expression
  "information retrieval"
- Proximity operators
  information NEAR retrieval

Evaluation

- Precision
- Recall
- Precision/recall graphs
- 11 point average precision
- TREC (Text Retrieval Conference)
- TREC Tasks: ad-hoc, web, spoken documents, multimedia, cross-language ...

Relevance

- Relevance is matching of a document with an information need expressed through a query
- Relevance is considered as binary and determined by human assessors for document-query pairs
- Relevance is modelled by a similarity measure that compares query and document representations

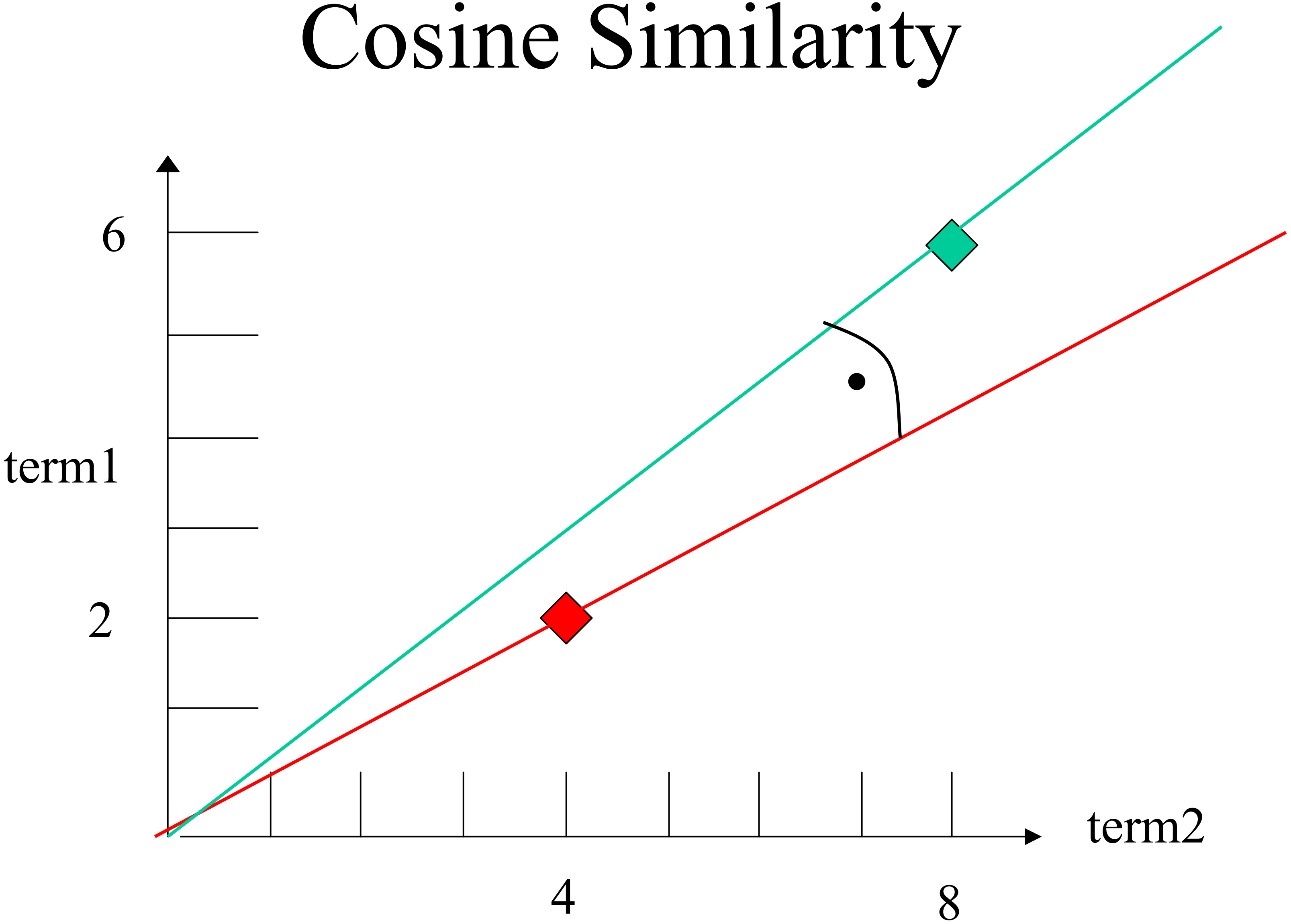
Document and Query Representations in IR

- Documents and queries are generally represented as a vector of terms weights
- Documents are treated as bags of words
- Preprocessing: stemming or morphological analysis
- POS, chunking, syntax did not improve information retrieval performance

Similarity Measures

- Term weighting: TF , TF/ICF, TF/IDF
- Similarity measures determine how close two documents are, or how alike a document and a query are
- A common similarity measure is the cosine of the angles between the vector representations

Cosine Similarity



Result Ranking

- Adjacency or proximity of search terms can be taken into account in ranking of retrieval results
- This accounts for phrases and collocations
- Search terms occurring near each other (e.g. within a paragraph) are more likely to be related than search term occurring in different parts of a document

Latent Semantic Indexing

- LSI: Singular value decomposition, dimensionality reduction
- LSI associates terms that share the same context, i.e. can be substituted
- Applications: information retrieval, cross-language IR, language learning, text categorisation, vocabulary tests

Query Expansion

- Query expansion with related terms (e.g. from WordNet, thesarus)
- Relevance Feedback: Query expansion with terms from relevant document
- Blind Relevance Feedback: Query expansion with terms from top-ranking document. Expansion with co-occurring terms improves precision/recall.

Language Models for IR

- Language models generate queries from documents
- Estimate probability that a given query was generated by a particular document
- Uni-gram language models
- (special case of probabilistic IR)

Cross-language IR

- Methods: document translation, query translation, parallel/comparable corpora

Document Categorization

- Similar techniques to IR (document representation, similarity measures)
- Document base contains categorized documents
- New document as query which retrieves the best matching documents from database
- Support Vector Machines achieve very good performance on various text categorization tasks

Document Clustering

- Similar techniques to IR (document representation, similarity measures)
- Each cluster is represented by a centroid
- Iterative hierarchical grouping of similar documents

Summarization

- Two approaches: Extraction (of sentencs or keywords) and abstraction (summary generation)
- Indicative vs. informative summaries
- Query-independent vs. query-biased summaries
- Evaluation criteria: informativeness, coherence

Information Extraction

- Tasks: named entity extraction, coreference, template extraction
- named entities: person, organisation, location, time, date, money, percentage
- methods: finite-state grammars, finite-state transducers
- evaluation: precision, recall, f-measure

Question Answering

- Answer extraction (passage retrieval) vs. Information extraction + answer generation
- Combination of IR-based and NLP-based approaches (semantic concepts, dependency relations).
- TREC open domain QA evaluation: extract 50-word passage containing the answer to a factual question

Co-location spaces

- Linear (speech, text)
- document (as bag of words)
- hierarchical structure (tree, dependency relations)
- semantic/conceptual space (e.g. WordNet)
- cyberspace (hyperlinks)

Collocations and IM

Collocations are

- multi-word units
- with statistical associations
- with restricted semantic compositionality

Are they useful for information management applications?

Collocations and Document Representations

- Common representations treat terms in the document and query as independent
- Collocations research shows that they are not independent
- Implications?

Collocations / Query Formulation

- Use of collocations for query expansion?

  (e.g. *collocation, corpus, association ...*

  vs. *collocation, facility, service, server, hosting ...)*

- Automatic or interactive?

Collocations / Categorisation & Clustering

- How much can category-specific collocations improve performance?
- Collocations for identification of genre, author, dialect?

Collocations / IE

- IE techniques (finite-state shallow parsing) for collocation identification
- Use of collocation in IE grammars (Gewinn machen, Umsatz erzielen ...)

Collocations / QA

- Use of collocations for finding answers (e.g. function-proper_name)

Collocations and Summarisation

- Keyword / key phrase extraction
- Evaluation of coherence: which association measures can be used?

Questions for Discussion

- Are collocations a useful level of representation for indexing and retrieval?
- Or are they only useful in establishing semantic representations?