Information Retrieval Methods

Helena Ahonen-Myka

Spring 2007, part 3

Indexing (1/2), presentation of the project work

Translation from Finnish: Greger Lindén

In this part

- Indexing
  - Different kinds of document descriptors
  - Goals of indexing
  - The influence of term frequency on the choice of terms → term weights
  - The influence of term discrimination on the choice of terms
  - Presentation of the project work

Indexing

- Construction of document descriptors
  - Usually by selecting a set of terms that are included in the description of the document
- Also constructing an index (= implementation of an index data structure) is part of indexing
- An index = a directory with
  - a set of document descriptors
  - search data structure
- Descriptors for the queries are constructed in mainly the same way as for documents

Description types

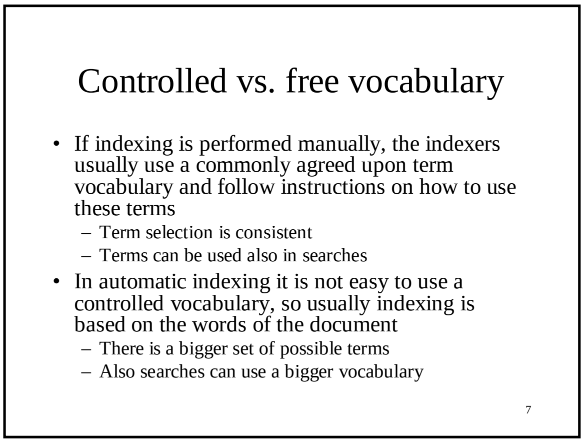
- Terms can be selected automatically or manually
- Terms can be objective or subjective
- Terms can be chosen from a controlled vocabulary or freely (usually from the text of the document)
- Terms can be single words or phrases (a word and its context)

Manually vs. automatically

- Before indexing was made manually
  - by experts in the field
  - or by indexing experts
- Today indexing is usually made automatically
  - Automatic indexing is never perfect, but also human indexers make mistakes or act illogically

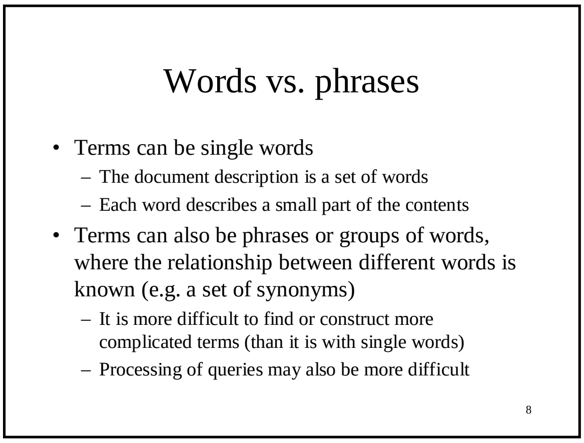
Objective vs. subjective terms

- Objective terms
  - Author of a publication, publication place, number of pages, and other bibliographic information
- Subjective terms
  - Terms that describe the contents of the document



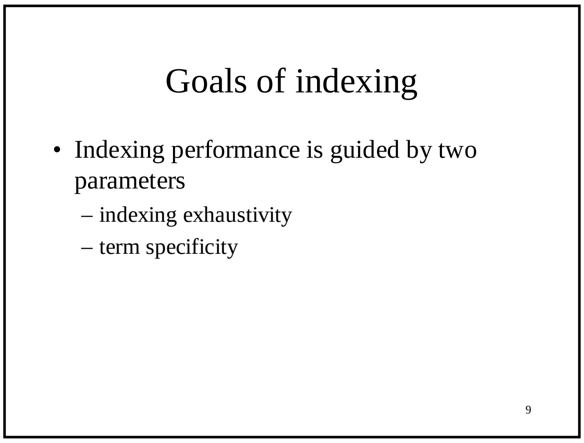
Controlled vs. free vocabulary

- If indexing is performed manually, the indexers usually use a commonly agreed upon term vocabulary and follow instructions on how to use these terms
  - Term selection is consistent
  - Terms can be used also in searches
- In automatic indexing it is not easy to use a controlled vocabulary, so usually indexing is based on the words of the document
  - There is a bigger set of possible terms
  - Also searches can use a bigger vocabulary



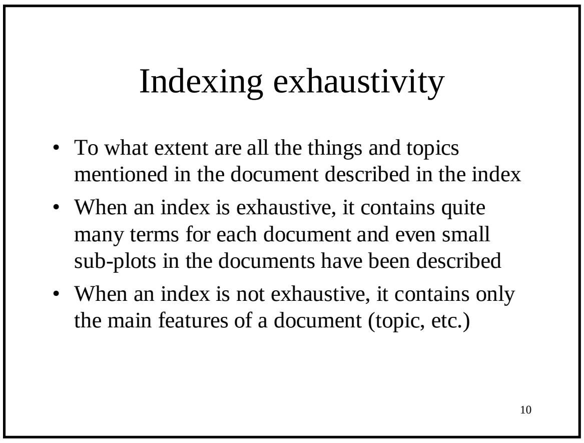
Words vs. phrases

- Terms can be single words
  - The document description is a set of words
  - Each word describes a small part of the contents
- Terms can also be phrases or groups of words, where the relationship between different words is known (e.g. a set of synonyms)
  - It is more difficult to find or construct more complicated terms (than it is with single words)
  - Processing of queries may also be more difficult



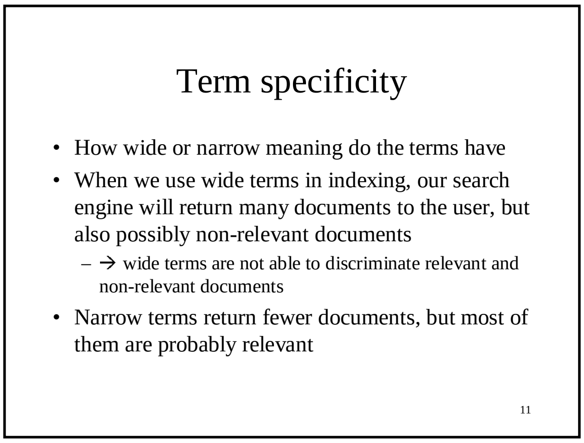
Goals of indexing

- Indexing performance is guided by two parameters
  - indexing exhaustivity
  - term specificity



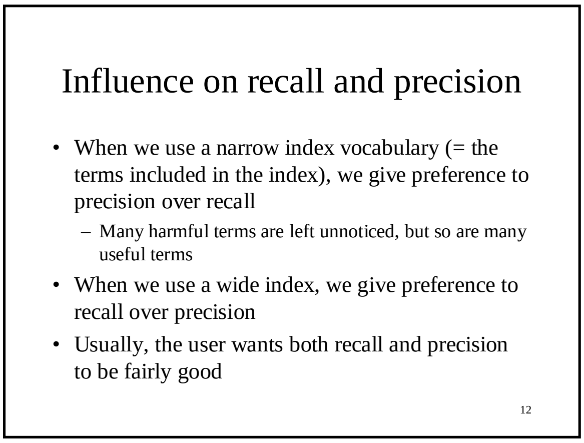
Indexing exhaustivity

- To what extent are all the things and topics mentioned in the document described in the index
- When an index is exhaustive, it contains quite many terms for each document and even small sub-plots in the documents have been described
- When an index is not exhaustive, it contains only the main features of a document (topic, etc.)



Term specificity

- How wide or narrow meaning do the terms have
- When we use wide terms in indexing, our search engine will return many documents to the user, but also possibly non-relevant documents
  - → wide terms are not able to discriminate relevant and non-relevant documents
- Narrow terms return fewer documents, but most of them are probably relevant



Influence on recall and precision

- When we use a narrow index vocabulary (= the terms included in the index), we give preference to precision over recall
  - Many harmful terms are left unnoticed, but so are many useful terms
- When we use a wide index, we give preference to recall over precision
- Usually, the user wants both recall and precision to be fairly good

Influence on recall and precision

- Exhaustive indexing may influence both recall and precision in the same direction
- If indexing is not exhaustive, the recall suffers: if the entity of interest has not been described at all in the index, the document will not be found
- The precision may suffer if some terms selected to the index are wide and do not discriminate relevant and non-relevant documents
  - If many terms are selected, it is more probable that some terms are wide

Term frequency as a base for term selection

- Our goal is to select terms that discriminate relevant and non-relevant documents
- But when we create the index we do not yet know what kind of searches we are going to use
- So we do not know what makes a term relevant or non-relevant
- But we can study the frequency of a term in the document collection and use this information

Term frequency as a base for term selection

- There are some words that occur frequently and evenly in every document
  - In English: and, the, in, of, ..
  - In Finnish: ja, ei, on, se, että, ...
  - In Swedish: och, den det, en ett, men, ...
- These words have a functional role, but they do not generally describe the contents of the document
- These words are collected in a stopword list (hukkasanalista)

Term frequency as a base for term selection

- Other words than the stopwords describe the document better
- These other words are usually not evenly distributed in the collection
- We can use their frequency when selecting terms for the index
- Idea: if a term occurs frequently in a document, it describes the content of the document

Term frequency as a base for term selection

- A possible indexing method
  - Remove all stopwords from the documents
  - Compute the term frequency tf_{ij} for all remaining terms T_j in each document D_i:
    - tf_{ij} = number of occurrences of term T_j in document D_i
  - Choose threshold K for the frequency and insert in the description of each document D_i the terms T_j, for which $tf_{ij} > K$

Term frequency as a base for term selection

- If we take into account only the number of occurrences of a term in the document, we will give preference to recall over precision
- Let’s assume that the term ”apple” occurs more than K times in each document in a set of documents
  - These documents surely tell about apples
  - If the user’s query contains the term ”apple”, these documents will be retrieved easier if ”apple” has been chosen as a term in the index
- What if the collection contains only documents describing apple growing?

Inverse document frequency

- The precision of the result will increase if the index contains terms that only occur in a small subset of the document collection
  - These terms discriminate efficiently this smaller subset from the rest of the documents
- Let the document frequency df_j be the number of documents where the term T_j occurs (at least once)

Inverse document frequency

- The discrimination ability of a term is described by the inverse document frequency, idf
- idf can be computed in many ways
- A usual way

$$idf_j = \log \frac{N}{df_j}$$

- Where N is the total number of documents in the collection

Term weight in a document

- Both recall and precision can be increased if we take into account the term frequency in a document (tf) and the term distribution in all the documents (idf)
- For the description, we should choose terms that occur frequently in the document in question but seldom in other documents
- We can define weight w_{ij} of term T_j in document D_i

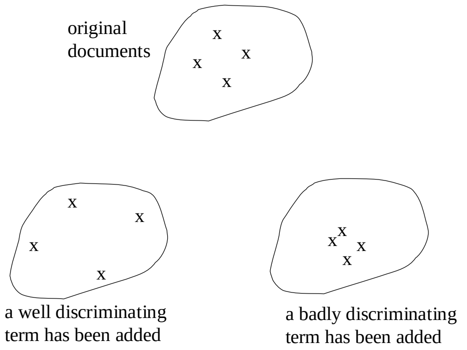
$$w_{ij} = tf_{ij} \cdot \log \frac{N}{df_j} = tf_{ij} \cdot idf_j$$

Term weight in a document

- A better indexing method
  - Choose a threshold K’
  - Remove all stopwords from a document
  - Compute the tf·idf weight w_{ij} for all terms T_j in all documents D_i
  - Select those terms j for the document i’s description whose weight w_{ij} exceeds the threshold K’

Term discrimination ability as a base for term selection

- A good property of the terms is that they discriminate well
- In the following slide each x denotes a document and the distance between the x’s denote how different the documents are (based on their terms)
  - x’s that are closer are more similar than x’s that are farther from each other
- The discrimination ability of a term can be measured: we compare the situation when the term has been selected to describe the document to the situation when it has not


original documents
a well discriminating term has been added
a badly discriminating term has been added

The discrimination ability of a term

- When we add a term that discriminates badly, e.g., a very common word, the document descriptions will become more similar
  - The average distance between documents will decrease (and the density of the collection will increase)
- When we add a term that discriminates well, usually a quite rare term, the documents where this term occurs will be more different than the others
  - The average distance between documents will increase (and the density of the collection will decrease)

The discrimination ability of a term

- The discrimination value dv_j of term T_j is computed as the change in the density of the document collection when the term is added to the descriptions of the documents

$$dv_j = Q - Q_j$$

- Q: density before the term is added
- Q_j: density after the term has been added

Density of the collection

- The density of a collection can be computed by the following expression

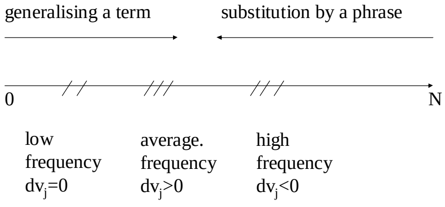
$$Q = \frac{1}{N(N-1)} \sum_{i=1}^{N} \sum_{\substack{k=1 \\ i \neq k}}^{N} sim(D_i, D_k)$$

- We compute the mutual similarity between all document pairs
- sim() is a similarity function that is based on the document terms

Discrimination ability vs. document frequency of terms

- A term that discriminates well gets a positive discrimination value
  - The average similarity between documents is lower when the term has been added
  - Such terms occur quite frequently on average in the collection
- Frequently occurring terms get a negative discrimination value
- Very rare terms do not influence the mutual similarity much → the discrimination value is close to zero

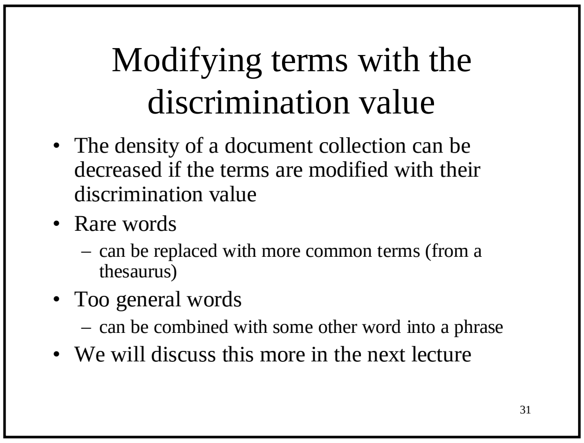
Discrimination ability vs. document frequency of terms?



Discrimination ability of a term vs. idf

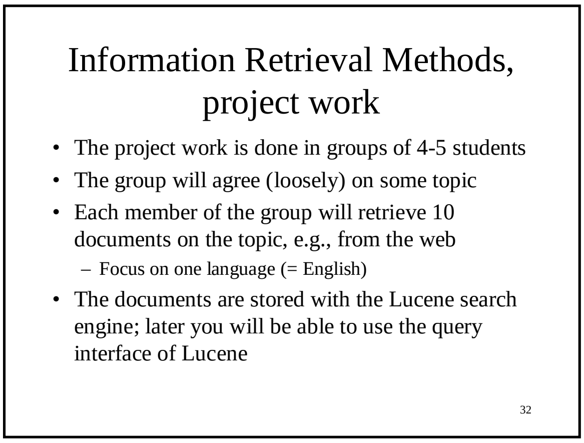
- The idf of a term decreases when the term frequency increases
- The discrimination ability increases when the frequency increases (low frequency → average frequency), but decreases when the frequency increases even more (average → high)
- If the term is weighted with the discrimination value instead, it will behave differently

$$w_{ij} = tf_{ij} \cdot dv_j$$



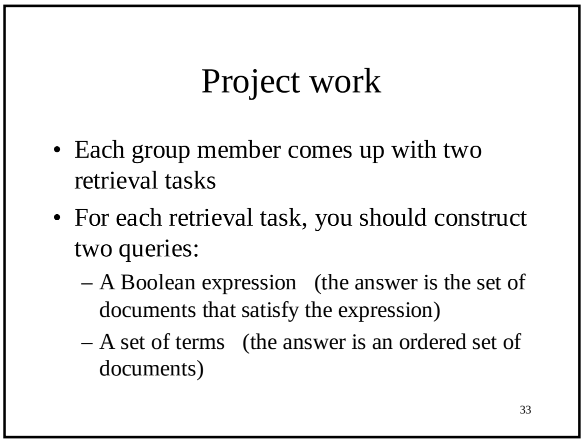
Modifying terms with the discrimination value

- The density of a document collection can be decreased if the terms are modified with their discrimination value
- Rare words
  - can be replaced with more common terms (from a thesaurus)
- Too general words
  - can be combined with some other word into a phrase
- We will discuss this more in the next lecture



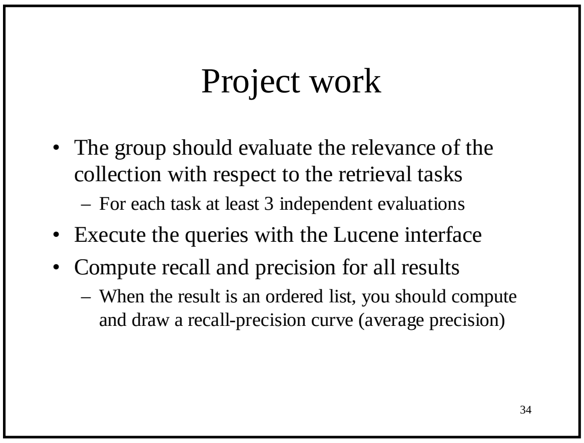
Information Retrieval Methods, project work

- The project work is done in groups of 4-5 students
- The group will agree (loosely) on some topic
- Each member of the group will retrieve 10 documents on the topic, e.g., from the web
  - Focus on one language (= English)
- The documents are stored with the Lucene search engine; later you will be able to use the query interface of Lucene



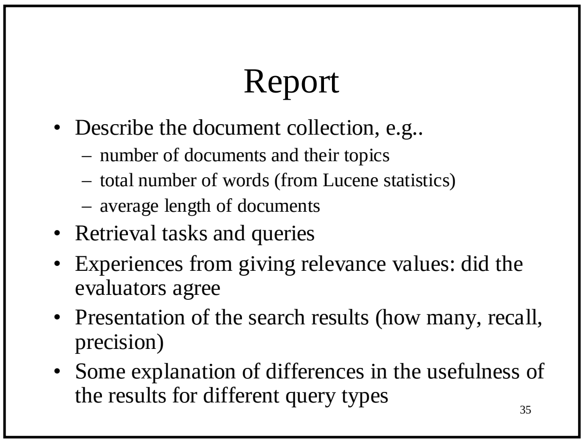
Project work

- Each group member comes up with two retrieval tasks
- For each retrieval task, you should construct two queries:
  - A Boolean expression (the answer is the set of documents that satisfy the expression)
  - A set of terms (the answer is an ordered set of documents)



Project work

- The group should evaluate the relevance of the collection with respect to the retrieval tasks
  - For each task at least 3 independent evaluations
- Execute the queries with the Lucene interface
- Compute recall and precision for all results
  - When the result is an ordered list, you should compute and draw a recall-precision curve (average precision)



Report

- Describe the document collection, e.g..
  - number of documents and their topics
  - total number of words (from Lucene statistics)
  - average length of documents
- Retrieval tasks and queries
- Experiences from giving relevance values: did the evaluators agree
- Presentation of the search results (how many, recall, precision)
- Some explanation of differences in the usefulness of the results for different query types