Information Retrieval Methods

Helena Ahonen-Myka
Spring 2007, part 10
Approximate matching: n-grams
From information need to query
Translation from Finnish: Greger Lindén

In this part

- Approximate matching
  - n-grams
  - s-grams
- From information need to query
  - Levels of storage and retrieval
  - Conceptual analysis of the retrieval task → retrieval plan
  - Modifying retrieval plans and queries

Approximate matching

- In the previous parts of the course (indexing, matching with the vector model, text-scanning) we have searched for whole words or word stems
  - e.g. the search word matches the index term when the strings are identical
- But there are also situations when we cannot be sure about the correct spelling of the word or the stem

Approximate matching

- Search words and index terms may be misspelled
- Foreign words, especially names, may be written in varying ways
  - e.g. Peking, Beijing
  - automatic morphological analyzers do not always find a correct base form for these words → each inflection (in index, in query) is a different string

Cross-language information retrieval (CLIR)

- Query in one language
- Answer documents in a different language (or many different languages)
- Dictionary-based CLIR
  - A bilingual dictionary is used to translate query words
  - Problem: query words are often proper names and technical terms that are not included in dictionaries
  - Often original terms are then used
  - Better solution: approximate matching
  - Proper names and technical terms are often close variants in different languages

Approximate matching techniques

- text scanning (string matching) methods have approximate matching variants
- Soundex, Phonix
  - phonetic codes are computed for the strings that are compared
  - the strings with similar codes are counted similar
- n-gram methods
  - language-independent

Approximate matching: n-grams

- Words can also be matched with approximate methods, with e.g., the n-gram method
- n-gram: a word’s substring of length n
  - usually n= 2 (bigram, sometimes digram) or n=3 (trigram)
  - usually sequential characters
- ”computer”
  - bigrams: *c, co, mp, pu, ut, te, er, r*
  - trigrams: *co, com, omp, mpu, put, ute, ter, er*

Similarity of a set of n-grams

- ”computer”
  - N1: {*c, co, om, mp, pu, ut, te, er, r*}
- ”compuetr”
  - N2: {*c, co, om, mp, pu, ue, et, tr, r*}
- We can compute the similarity with the measure

$$SIM(N_1, N_2) = \frac{|N_1 \cap N_2|}{|N_1 \cup N_2|}$$

- 6/12 = 0.50

Similarity of a set of n-grams

- trigrams:
  - computer: *co, com, omp, mpu, put, ute, ter, er*
  - compuetr: *co, com, omp, mpu, pue, uet, etr, tr*
- 4/12 = 0.3

Skip grams (s-grams)

- The characters in n-grams do not have to be sequential
- Pirkola et al introduce s-grams (skip grams)
  - focus is on bigrams (Note: the article uses “digram”)
  - let’s assume that we are looking at the word w
  - CCI (character combination index) denotes the number of skipped characters
  - e.g., CCI=(2) denotes s-grams constructed from a word w, where the s-bigram characters are within two characters from each other

An example of s-bigram classes

word	CCI	s-bigrams
pharmacology	(0)	{ph, ha, ar, rm, ma, ac, co, ol, ...}
	(1)	{pa, hr, am, ra, mc, ao, cl, oo,...}
	(2)	{pr, hm, aa, rc, mo, al, co, og, ly}
farmakologian	(0)	{fa, ar,rm, ma, ak, ko, ol, lo,...}
	(1)	{fr, am, ra, mk, ao, kl, oo, lg,...}
	(2)	{fm, aa, rk, mo, al, ko, og, li, oa, gn}

S-bigrams

- The CCI and the combinations of bigrams affect how well the matching will succeed
- If we use all bigrams (CCI = (0,1,2,...,m-2)), where m is the length of the word, there are probably very common bigrams included that occur in many words
- If we only use sequential characters, even many similar words will not match

S-bigram classes

- We can make the matching more precise by using s-bigram classes
- Unclassified bigrams
  - CCI is of the form (i, i+1,…,i+j), i, j ≥0
- Using classes in matching
  - We define classes in the CCI, e.g. ([0], [1,2])
    - class [0] and class [1,2]: the set of adjacent bigrams and the set of bigrams with a skip of 1 or 2 characters

Examples of unclassified and classified s-bigrams

word	CCI	s-bigrams
abcde	(0)	{ab,bc,cd,de}
	(0,1)	{ab,ac,bc,bd,cd,ce,de}
	(0,1,2)	{ab,ac,ad,bc,bd,be,cd,ce,de}
	([0],[1])	{ab,bc,cd,de} {ac,bd,ce}
	([0],[1,2])	{ab,bc,cd,de} {ac,ad,bd,be,ce}
abce	(0,1)	{ab,ac,bc,be,ce}
	([0],[1,2])	{ab,bc,ce}{ac,ae,be}

Matching of s-grams

- Unclassified bigrams
  - each bigram according to the CCI of word w1 is compared to each bigram according to the CCI of word w2
  - abcde vs. abce: CCI(0,1): 4/10 = 0.4
- Classified bigrams
  - we only try matching bigrams of the same classes
  - abcde vs. abce: CCI([0], [1,2]): max(2/5,2/6) =0.4

Results

- The study compared medical terms and place-names in English, German and Swedish to corresponding terms and names in Finnish
- Classified s-bigrams gave a better result than using n-grams (where no skips are allowed)
- Also unclassified s-bigrams work better than bigrams
- Especially when words are short, the s-bigrams are better
  - But when the words are very short, no method works well

Possible improvements

- S-gram frequencies could be taken into account
  - high-frequency s-grams do not discriminate well → down-weighting of these terms
- Translitteration rules could be used
  - sometimes variation is too high
    - Chechnya (English) vs. Tsetshenia (Finnish)
  - translitteration rules for different language pairs could be generated automatically

From information need to query

- Levels of storage and retrieval
- Conceptual analysis of the retrieval task → retrieval plan
- Modifying retrieval plans and queries

Levels of storage and retrieval

- Retrieval tasks and documents can be represented on three levels
  - – conceptual level
  - – expression level
  - – occurrence level

Conceptual level

- On the conceptual level we focus on the concepts and concept relations of the retrieval task and the documents
- We have to take into account
  - – the goal of the retrieval task: just a few documents vs. everything about the topic
  - – the concepts used in the relevant documents in the document collection
    - even concepts in relevant documents can differ
- “at best” analysis by a human, but today the conceptual phase is not usually considered in indexing or retrieval

Expression level

- The expression level denotes the ways of expressing concepts in a natural language (or some specialised language)
- The representation of a document on the expression level is the document itself + possibly some other terms
- Task: to find alternative expressions for concepts

Expression level

- From conceptual level to expression level:
  - – synonyms, phrases
  - – variants in standard language, business language, scientific language
  - – current, recommended, or to-be-avoided (obsolete) terms
  - – spelling variants
  - – abbreviations and full denominations
  - – divided/combined variants (“data base”, “database”)
  - – more specific or more general terms

Occurrence level

- Concrete retrieval always happens at the occurrence level (character level)
  - – basic operations: similarity of strings, relations between the positions of occurrences of strings, number of occurrences, etc.
- When forming the query, key words on the expression level are converted into strings of the occurrence level
  - – e.g. concatenations, stemming, regular expressions (wild cards), n-grams
- The user may skip the conceptual and expression levels and give the query directly

Conceptual analysis of the retrieval task

- The meaning of the conceptual analysis is to recognise the central concepts of the search topic and the relationships between the concepts
- The result is a conceptual retrieval plan, based on which the query is implemented
  - – conceptual retrieval plan: what information should be retrieved
  - – query: how the information is retrieved
- The conceptual analysis makes it easier to recognise the relationships between necessary concepts
  - – focusing on a few concepts is easier to manage than processing a large set of search words

Conceptual analysis of the retrieval task

- Concepts can have parallel and restricting relationships
  - parallel relationships: the concepts represent the same facet (aspect, point of view) of the search topic
  - restricting relationships: the concepts represent different facets
- The relationships are often association relationships
  - actor – act – action – tool
  - act – object – result
  - reason – consequence

Conceptual analysis of the retrieval task

- Only because two concepts are associated, we cannot say that they belong to the same facet
- The solution depends on how the user combines concepts
  - "productivity and support for forestry and stock breeding"
    - "productivity" and "support" represent the same facet
  - "the influence of forestry and stock breeding support on the productivity during the 1980s"
    - "productivity" and "support" represent different facets

Conceptual analysis of the retrieval task

- Parallel relationships: expressions representing the concepts are combined with a disjunction (OR) in a Boolean query
- Restricting relationships: we use conjunctions (AND) and negations (NOT)
- ”pollution produced by fish”
  - pollution AND fish NOT (fishing industry AND water protection)

Completeness, accuracy and coverage of a retrieval plan

- A retrieval plan has three types of characteristics
  - completeness
  - accuracy
  - coverage

Completeness, accuracy and coverage of a retrieval plan

- Completeness
  - How many of the facets of the search topic are included in the retrieval plan
  - A facet is included in the plan, if the plan contains some concept that belongs to the facet
  - Completeness always concerns restricting relationships of the facets of the search topic

Completeness, accuracy and coverage of a retrieval plan

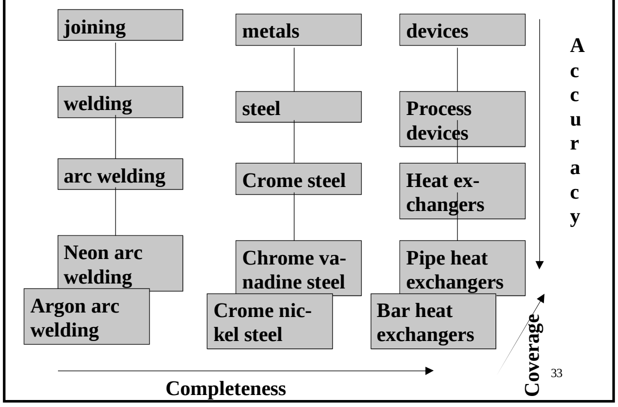
- Accuracy
  - How accurate are the concepts that describe the facets in the retrieval plan
  - The retrieval plan is completely accurate if the facets are represented exactly on the level of accuracy of the search topic; otherwise it is not accurate
  - Accuracy is always concerned with hierarchical relations between the concepts

Completeness, accuracy and coverage of a retrieval plan

- Coverage
  - How many concepts are used to describe each facet in the retrieval plan
  - A retrieval plan has coverage if it contains all dimensions of each facet; otherwise it does not have coverage
  - Coverage always concerns internal parallel relationships within a facet of a search topic

Completeness, accuracy and coverage of a retrieval plan

- On the following slide: three facets whose top concepts according to the conceptual hierarchy are “joining”, “metals” and “devices”
- We assume that the description of the search topic is described exactly by the concepts “Argon arc welding”, “Neon arc welding”, “chrome nickel steel”, “pipe heat exchangers” and “bar heat exchangers”
  - “Argon arc welding” and “neon arc welding” are parallel concepts
  - “Pipe heat exchangers” and “bar heat exchangers” are parallel concepts



Completeness, accuracy and coverage of a retrieval plan

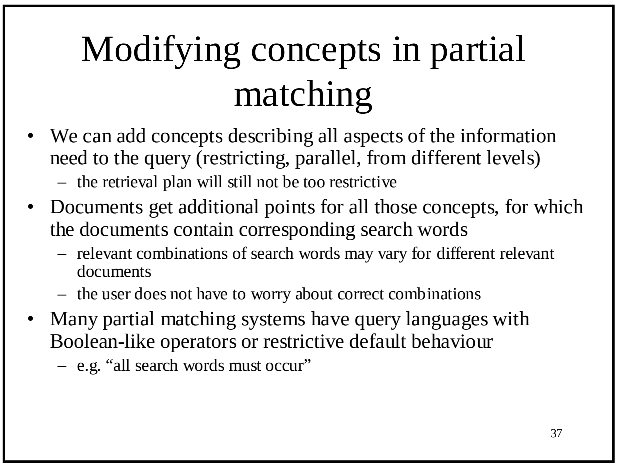
- The retrieval plan is absolutely complete if it contains every facet of the information need
  - e.g.. “arc welding”, “steel” and “heat exchangers”
- If the retrieval plan uses the concepts “Argon arc welding” and “bar heat exchangers”, the plan is accurate in relation to these facets
  - if we use concepts like “joining” and “devices”, the plan is not accurate (in relation to these facets)
- The plan has coverage in relation to the “devices” facet if it contains the concepts “pipe heat exchangers” and “bar heat exchangers”

Modifying concepts in exact matching

- If we increase the completeness of a retrieval plan, the search will be narrower
  - usually means adding restricting concepts
  - the recall decreases, the precision increases and the size of the result decreases
- Increasing the accuracy will also make the search narrower
  - the recall decreases, the precision increases and the size of the result decreases
- Increasing coverage will widen the retrieval
  - means adding parallel concepts related to a certain facet
  - the recall increases, the precision decreases and the size of the result increases

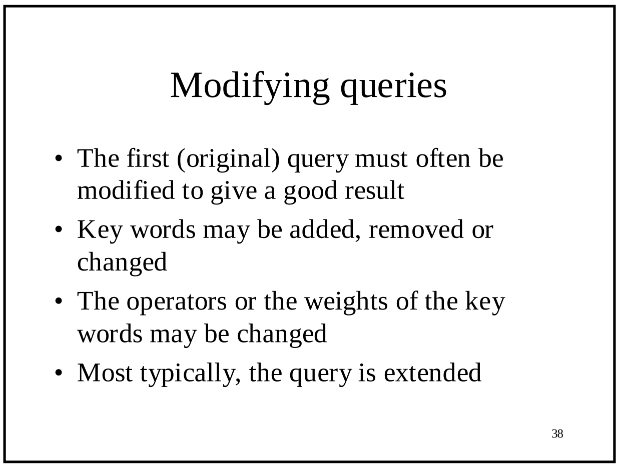
Modifying concepts in partial matching

- When we use partial matching, there are no parallel or restricting relationships for the concepts
- If we leave out an important concept from the retrieval plan → the documents will not get “additional points” for containing terms corresponding to the concept
  - a significant document may stay under the threshold for being retrieved



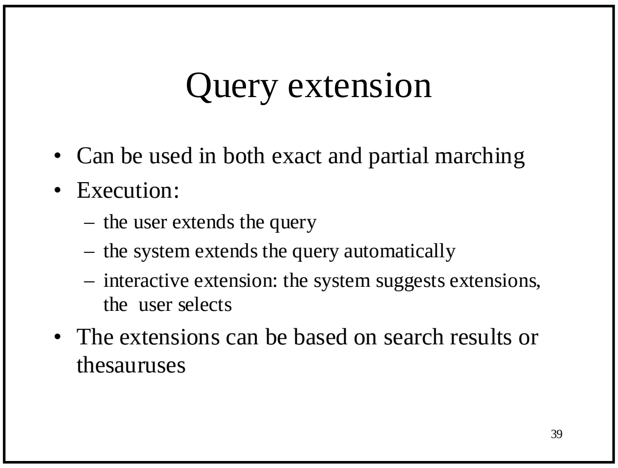
Modifying concepts in partial matching

- We can add concepts describing all aspects of the information need to the query (restricting, parallel, from different levels)
  - the retrieval plan will still not be too restrictive
- Documents get additional points for all those concepts, for which the documents contain corresponding search words
  - relevant combinations of search words may vary for different relevant documents
  - the user does not have to worry about correct combinations
- Many partial matching systems have query languages with Boolean-like operators or restrictive default behaviour
  - e.g. “all search words must occur”



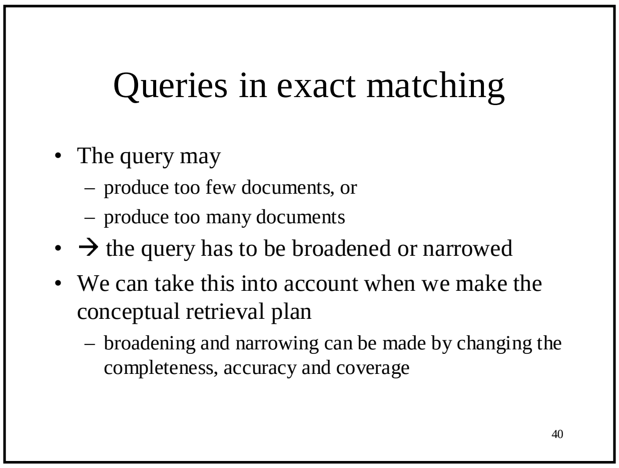
Modifying queries

- The first (original) query must often be modified to give a good result
- Key words may be added, removed or changed
- The operators or the weights of the key words may be changed
- Most typically, the query is extended



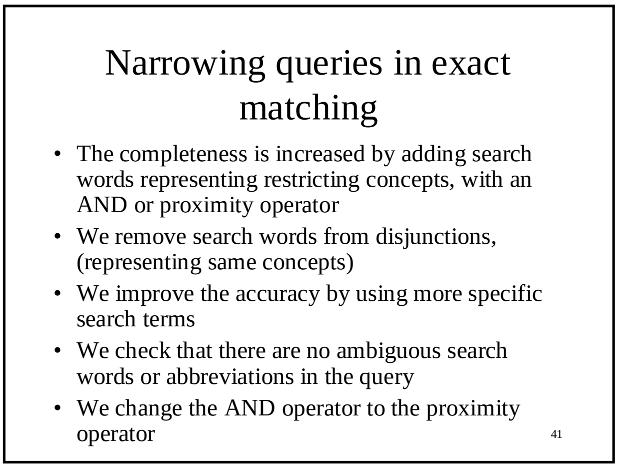
Query extension

- Can be used in both exact and partial marching
- Execution:
  - the user extends the query
  - the system extends the query automatically
  - interactive extension: the system suggests extensions, the user selects
- The extensions can be based on search results or thesauruses



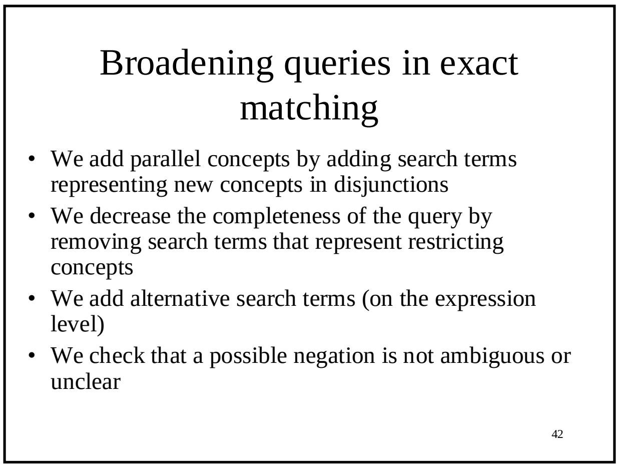
Queries in exact matching

- The query may
  - produce too few documents, or
  - produce too many documents
- → the query has to be broadened or narrowed
- We can take this into account when we make the conceptual retrieval plan
  - broadening and narrowing can be made by changing the completeness, accuracy and coverage



Narrowing queries in exact matching

- The completeness is increased by adding search words representing restricting concepts, with an AND or proximity operator
- We remove search words from disjunctions, (representing same concepts)
- We improve the accuracy by using more specific search terms
- We check that there are no ambiguous search words or abbreviations in the query
- We change the AND operator to the proximity operator



Broadening queries in exact matching

- We add parallel concepts by adding search terms representing new concepts in disjunctions
- We decrease the completeness of the query by removing search terms that represent restricting concepts
- We add alternative search terms (on the expression level)
- We check that a possible negation is not ambiguous or unclear

Broadening queries in exact matching

- We decrease the accuracy by removing too specific search terms (and replacing them with more general)
- We replace the search terms with regular expressions which match a greater number of words (e.g. use of wild cards)
- We change proximity operators to AND operators

43

Queries in partial matching

- The above tricks do not work with queries used in partial matching
  - adding search terms in a query does not decrease the size of the answer set but it may increase precision
- Typical modifications: adding, removing or giving different weights to search terms based on relevance feedback

44

Queries in partial matching

- Automatic extension of a query can be based on
  - Independent word lists and concept hierarchies (independent of the collection, e.g., thesauruses)
  - Word lists based on the associations of words in the collection
- There are somewhat conflicting results about how effective these extensions are
  - Longer queries produce better results than shorter ones → extensions pay off
  - Added search words must occur in the collection → external collections do not necessarily produce good results

45

In this part

- Approximate matching : n-grams, s-grams
- From information need to query
  - Levels of storage and retrieval: conceptual level, expression level, occurrence level
  - Conceptual analysis of the retrieval task
  - Completeness, accuracy and coverage of the retrieval plan
  - Concepts in exact and partial matching
  - Queries in exact and partial matching

46