LBSC 878 Oard/**Soergel** Sp 1999

Dagobert Soergel ds52@umail.umd.edu March 7, 1999

Week 6. March 8, 1999

The matching function in document retrieval

Some of this applies to retrieval generally

This is a work in progress. The framework will be extended to cover feature assignment and selection/examination/use as well. Given the unity of the entire retrieval process, where the same subprocess might occur in feature assignment in one system and in matching in the other, such an extension makes more sense.

There is no claim for completeness; any additions are highly welcome. A data flow diagram might be even more useful or at least complementary to the table format.

A further caveat: This overview considers only retrieval methods that judge each document on its own, without considering the total retrieval result. But often it is important to look at the total result.

> Example 1. Building a "staircase" from the user's present knowledge to the desired knowledge; documents on one step help the user understand the documents on the next step.
>
> Example 2. In litigation support systems it is often important to find a group of documents that demonstrate a pattern of behavior.

Systems producing this kind of result are very challenging to build.

The matching function is analyzed in terms of four ingredients:

1 Information/knowledge used

2 Use of this information in matching — Matching methods and processes

3 How to represent this information (given only for a few of the types of information)

4 Where to get this information

Outlines for 1 and 2 are presented first, followed by a table that shows for each type of information the method or process it supports as well as information on 3 and 4.

Note: NLP means Natural Language Processing

Information/knowledge used in matching

Outline

Information about documents

Information about requirements. User model

"External" information/knowledge

Information about documents

Note: Some of these types of information are themselves derived from more basic types as discussed in the outline on feature assignment..
The kinds of information are arranged roughly in a dependency order (like a food chain), the most basic ones first and those that are derived later.

The full text of the document or the text of document sections (including title and abstract). (Automated processes assume text in machine-readable form)

Text of review or independent abstract of document

Typographic and document structure information. SGML or HTML codes in the document (even codes giving typographical information might be useful).

Analysis of the document structure, such as results of a parse or results of an automated segmentation of the document.

Document features with weights, for example, terms in the document with frequency of occurrence in the document, or concepts assigned automatically or manually with importance weights. Includes document type, subject domain, readability score, or other features.

Relationships among features within a document, such as

Statistical associations between two terms within the document

Semantic relationships, such as *A causes B* specified in the document or assigned to the document by an indexer

A frame with slots filles as a document representation

Relationships between documents, such as citation relationships. A citation relationship can be considered a feature

Context of a document

Information about requirements. User model

Explicit information on requirements

Free description of requirements (called *query statement* by some, *query formulation* by others)

A formal representation of requirements

Note: Ultimately, this is a structure that combines query features in accordance with the rules of the IR system (called *query formulation* by some, *query representation* by others). But it could also be a description that contains more or less explicitly all the information needed to build the query representation.

A formal representation of requirements should include

The features to be used or reasonably isolated concepts. This is an estimation as to which features would be good predictors of document relevance

The query weight for each feature. This is an estimation of the importance of this predictor in the matching formula

Interdependence among features. If feature A is present in a document, the predictive power of feature B for that document may be diminished. Extreme case: A OR B; if A is present, the presence of B contributes nothing (but if A is not present, B does contribute). The values here are estimates based on the meaning of the features for the query.

Interaction among features (in the statistical sense). The presence of A and B combined may contribute more to the relevance of a document than the sum of the individual contributions of A and B. Extreme case: A AND B; individually, A and B contribute nothing, only the presence of both does. The values here are estimates based on the meaning of the features for the query.

Information on user background

Implicit information about requirements

Relevance assessments of some documents (which might include relevant documents or irrelevant documents or both). These provide a learning corpus. . Various statistics can be derived from such a learning corpus.

Data on user behavior while perusing documents (How long did the user look at this document record, did she call up the document itself, did she print the document)

"External" information/knowledge

Note: Information/knowledge not directly about an individual document or information need

Knowledge about morphology and syntax (stemmers with exception dictionaries, part-of-speech-taggers, parsers)

Knowledge needed to disambiguate word senses (statistical associations, syntactic/semantic patterns)

Knowledge about document structure: how to parse a document or how to segment a document or how to interpret SGML or HTML codes.

General statistics about features, derived from corpora other than the IR system collection (features could be terms or concepts or ...), such as

- Relative frequency of a word or phrase or concept in a universal cross-section of texts or in a specific domain
- Relative frequency of the meanings of a homonym in a universal cross-section of texts or in a specific domain
- Term or concept co-occurrence data in a universal cross-section of texts or in a specific domain
- This information for different languages
- Relative frequency of the possible translations of a term in language A into language B, in a cross-section of texts or in a given domain

Corpus statistics about features (terms or concepts or ...)

- Total frequency of a term or concept in a corpus
- Number of documents in which a term or concept occurs
- Co-occurrence statistics, term or concept associations

Relationships between features independently from individual documents, such as

- Statistical associations (see above)
- Meaning relationships
  - Meaning relationships between terms (synonyms/quasisynonyms)
  - Meaning relationships between concepts (hierarchical, many types of associative relationships)

Knowledge about quality of sources (journals, publishers) and persons.

Matching methods/processes

Isolated feature matching

Matching methods producing two scores: 0 and 1

Boolean query formulation, possibly using weighted features in the query representations.

Note: Ranking can be achieved by running progressively broader Boolean queries

Matching methods producing a wider range of scores, which can be used for ranking

Vector space matching.

Probabilistic matching.

Inference networks.

Note: Inference networks were introduced as a method to implement probabilistic matching but can also be looked at as a general formalism that, with proper choice of parameters, can be used to represent Boolean matching and vector space matching.

Spreading activation in a semantic network.

Structure matching

Note: each of these could be further divided into methods that produce just two scores and methods that produce a wider range of scores and can thus be used for ranking.

Proximity matching

Distance-based proximity matching (within window of width x)

Structure-based proximity searching (within same sentence, same paragraph, etc.)

Matching syntactic structures in the natural language text.

Note: One can surmise that this would not work very well because the same meaning (deep structure) can be represented by many syntactic surface structures.

Matching structures in some knowledge representation

Frame-based retrieval. Matching on frame structure and not just specific slot values.

Matching structure in conceptual graphs (on a formal level, this is related to chemical structure searching).

Matching overall structure - indicator of document type, such as letter, will, scientific article

Retrieval using rules. Could couch any of the above in terms of rules.

Word sense disambiguation

All these methods, when applied in a free text environment, can be augmented by word sense disambiguation in the query or in the documents or both. Word sense disambiguation in the query is easy (just ask the user), word sense disambiguation in documents is hard. If query feature expansion is used, disambiguating query terms may be useful even if document terms are not disambiguated.

The effect of word sense disambiguation is highly dependent on query length. If the query is long, it is unlikely that a document will contain all the query terms in the unwanted meaning. So there is probably an implicit disambiguation effect.

query feature expansion

All these methods can be augmented through query feature expansion (usually referred to more narrowly as query term expansion) using document-specific or external knowledge about term relationships. The effect this may have on retrieval performance needs to be studied.

The effects of query term expansion depends on a number of factors, which must be considered when interpreting experimental results. Among them:

Are the query features concepts or words?

If the query features are concepts, expanding a query feature (adding more terms that designate the concept or, in hierarchic expansion, a narrower concept) means providing more ways in which that feature can match a document while maintaining the basic structure and length of the query. If the document has several of the added terms, the concept is still only matched once (possibly with increased weights)

If the query features are words or terms, adding more terms (synonyms or hyponyms) increases the length of the query. A document that has several synonymous terms will have a separate match for each, which may skew results.

Types of relationships used. Quality of relationships. Applied wholesale or selectively (for example, use all RT given in a thesaurus or only those that make sense for the query at hand)

Is word sense disambiguation used? (Not an issue with a controlled vocabulary)

Word sense disambiguation in the query? If not, each meaning of a query term will be expanded, adding many synonyms unrelated to the query topic.

Word sense disambiguation in the documents? If not, query term expansion multiplies the homonym problem: A term S added because one of its meaning is synonymous with one of the query term's meanings will add to the score of documents which use S in a different meaning.

Matching formula used

Query length

Kind of information	Used for	Represented as	Obtained from/how
Information about documents			
The full text of the document or the text of document sections (including title and abstract). (Automated processes assume text in machine-readable form)	Automated examination of documents found (e.g., for distance-based proximity searching, phrase searching)	String of bytes	Electronic version of text generated at the origin. OCR of pages. Speech recognition
Text of review or independent abstract of document	Further source of features, including quality ratings		
Typographic and document structure information. SGML or HTML codes in the document (even codes giving typographical information might be useful).	Structure-based proximity searching (within same sentence or para-graph). Weight of features assigned "on the fly". Matching syntactic structures in the natural language text.	Any of various schemes for document structure representation	Codes inserted at document creation.
Analysis of the document structure, such as results of a parse or results of an automated segmentation of the document.			Structure analysis based on typographical and layout information obtained from OCR or speech re-cognition. Segmentation of a document into sentences. Segmentation of a document into meaningful units. Parsing sentences
Document features with weights, for example, terms in the document with frequency of occurrence in the document, or concepts assigned automatically or manually with importance weights. Includes document type, subject domain, readability score, and other features.	Any of the matching methods	List of features, possibly group-ed into fields of a record	Manual indexing. Automated indexing
Information about documents, continued			

Kind of information	**Used for**	**Represented as**	**Obtained from/how**
Relationships among features within a document, such as	Word sense disambiguation on the fly.		
Statistical associations between two terms within the document			Possibly computed from occurrence in proximity windows or sentences
Semantic relationships, such as *A causes B* specified in the document or assigned to the document by an indexer	Structure matching	Any of several knowledge representation schemes	Manual indexing. Derived through NLP/semantic analysis
A frame with slots filles as a document representation	Frame-based matching	Frames	Manual indexing. Derived through NLP/semantic analysis
Relationships between documents, such as citation relationships. A citation relationship can be considered a feature	Any feature matching method can use cited or citing documents as features. Expanding the scope of features being considered by including features of cited or citing documents, possibly lowering their weights		From some database, such as Science Citation Index. Manual input. Automated extraction form text.
Context of a document	Any aspect of context can be used as a feature in matching		

Kind of information	Used for	Represented as	Obtained from/how
Information about requirements. User model			
Explicit information on requirements			
Free description of requirements. Might include such things as a project description, job description, or a persons CV and bibliography	Deriving query representation (manually, interactively, or automatically)	Text and/or list of words	User User with intermediary (reference interview)
Formal representation of requirements			User User dialog with the system User with intermediary Intermediary alone based on free description System based on free description System based on relevance judgments
The features to be used or reasonably isolated concepts. This is an estimation as to which features would be good predictors of document relevance	Matching algorithm		
The query weight for each feature. This is an estimation of the importance of this predictor in the matching formula	Matching algorithm		
Interdependence among features: If A, then B contributes less	Matching algorithm: If A is present, adjust weight of B		
Interaction among features (in the statistical sense). The presence of A and B combined contributes more to the relevance of a document than the sum of the individual contributions of A and B	Matching algorithm: If A and B are both present, adjust document score accordingly		

Kind of information	Used for	Represented as	Obtained from/how
Information about user requirements, continued			
Information on user background	Derive further query features, for example language of document or readability score		
Implicit information about requirements			
Relevance assessments of some documents (which might include relevant documents or irrelevant documents or both). These provide a learning corpus. Various statistics can be derived from such a learning corpus.	Adjusting query representation (any of the parameters) (relevance feedback)		Relevance assessments may come from the user, either explicitly or inferred form observation of user behavior, from other users, from experts, from citations in a known relevant document, or from the assumption that documents ranked highly in preliminary search are relevant
Reasons for relevance assessments	Allows for more focused adjustment of query parameters		User Inferred from relevance judgments
Data on user behavior while perusing documents	Estimating relevance assessments		From observation, generally through system logs

Kind of information	Used for	Represented as	Obtained from/how
"External" information/knowledge Information/knowledge not directly about an individual document or information need			
Knowledge about morphology and syntax (stemmers with exception dictionaries, part-of-speech-taggers, parsers)	Normally these would be applied in feature assignment, but they could be used for on-the-fly analysis, for example to check whether two words with a distance of 5 are syntactically related, which in turn might help with determining whether one of the words is used in the desired sense		Acquire existing tools Check The *Linguistic Data Consortium* (LDC, www.ldc.upenn.edu/ldc)
Knowledge needed to disambiguate word senses (statistical associations, syntactic/semantic patterns)			Acquire existing tools Check The *Linguistic Data Consortium* (LDC, www.ldc.upenn.edu/ldc
Knowledge about document structure: how to parse a document or how to segment a document or how to interpret SGML or HTML codes.	Determining document term weights based on position or typography. Structure-based proximity searching Matching based on overall structure		
General statistics about features, derived from corpora other than the IR system collection (features could be terms or concepts or ...), such as			
Relative frequency of a word or phrase or concept in a universal cross-section of texts or in a specific domain	Use as a substitute for document frequency in the corpus		
Relative frequency of the meanings of a homonym in a universal cross-section of texts or in a specific domain	"Best guess" homonym disambiguation. Eliminate obscure meanings as possibilities in a query		

Kind of information	Used for	Represented as	Obtained from/how
"External" information/knowledge, continued			
Term or concept co-occurrence data in a universal cross-section of texts or in a specific domain	Term dependence: If A is present, discount the weight of a highly correlated term B		
Relative frequency of the possible translations of a term in language A into language B, in a cross-section of texts or in a given domain	"Best guess" in cross-language information retrieval		
.Corpus statistics about features (terms or concepts or ...)			Computation based on access to a corpus
Total frequency of a term or concept in a corpus			
Number of documents in which a term or concept occurs	Any feature matching method as a component of computing weights in the ranking formula		
Co-occurrence statistics, term or concept associations	Word sense disambiguation Term dependency used in adjusting weights		
Relationships between features independently from individual documents, such as			
Statistical associations (see above)			
Meaning relationships			
Meaning relationships between terms (synonyms/quasisynonyms)	Query term expansion. Strength of relationship may be used to adjust the weight of the added terms	Traditional thesaurus Semantic nets Frames	Dictionaries, thesauri, classifications, ontologies
Meaning relationships between concepts (hierarchical, many types of associative relationships)			
Knowledge about quality of sources (journals, publishers) and persons.	Context		Reference works, reviews