

Onderwerpsontsluiting van Digitale Documenten

Jaap Kamps
University of Amsterdam

NVB-WN Studiedag Onderwerpsontsluiting
Koninklijke Bibliotheek, Den Haag, 27 april 2006

Outline: Subject Access to Digital Documents

- Importance of Subject Access
- Basics
 - ★ Information Retrieval
 - ★ Evaluation
- Automatic versus Manual Indexing
 - ★ Controlled Vocabularies
 - ★ Free text
 - ★ Results and Lessons
- Conclusions

Information Retrieval 101

- *Definition* of **Information Retrieval**:
 - ★ the discipline that deals with retrieval of unstructured data, esp. textual documents, in response to a query or topics statement.
- Key is the matching of
 - ★ user's information need
 - ★ with available information (usually documents)
- *Task* of an IR system is:
 - ★ Given a **document collection** and a **query**,
 - ★ return a **ranked list of documents** that are *relevant* for the query

IR 101: Sample Topic from TREC ad hoc

```
<top>
<num>303</num>
<title>Hubble Telescope Achievements</title>
<desc> Identify positive accomplishments of the Hubble telescope since it was launched in 1991.</desc>
<narr> Documents are relevant that show the Hubble telescope has produced new data, better quality data than previously available, data that has increased human knowledge of the universe, or data that has led to disproving previously existing theories or hypotheses. Documents limited to the shortcomings of the telescope would be irrelevant. Details of repairs or modifications to the telescope without reference to positive achievements would not be relevant.</narr>
</top>
```

Motivatie

- Digitale revolutie
 - ★ Praktisch alle nieuwe documenten zijn (ook) beschikbaar in digitale vorm
 - ★ Zelfs historische documenten worden op grote schaal gedigitaliseerd

- Centrale vraag:

Wat is het belang van **onderwerpsontsluiting** in de context van **digitale documenten**?

For Whom Do We Do It?

- Importance of "searching by subject"
 - ★ Does the Internet-generation need subject access points?
- Classification of Internet search engine queries Broder [2002]
 - ★ 50% is *informational* (↔ subject access points)
 - ★ 20% is *navigational* (↔ named access points)
 - ★ 30% is *transactional* (↔ ???)
- Answer seems **yes!**
 - ★ Internet users often search for **information**
 - ★ rather than a specific known site or document

IR 101: Evaluation

- System-based evaluation is abstraction of the retrieval process
 - ★ **Retrieval effectiveness**: Good system performance is equated with good document rankings
 - ★ Run as a **laboratory test** with control over variables affecting performance
 - ★ Thus allows for fair **comparative testing**
- **Test Collection** consisting of
 - ★ 1) **Topics**, 2) **Documents**, and 3) **Relevance judgments**
- Can be reused *ad infinitum*...

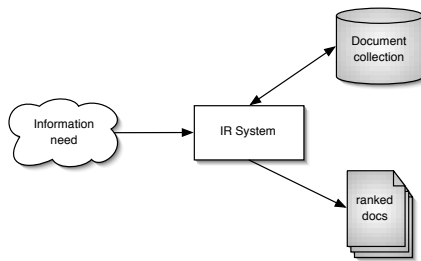
IR 101: Assessing Pooled Documents



- TREC uses a three judge system with majority vote.

General Picture of Document Retrieval

- document collection + query \rightsquigarrow ranked list of documents



Essence of Controlling Vocabulary

- Basic Idea** is to use a common language
 - to index documents and
 - to state search requests
- Matching takes place in this common language
 - which is free from ambiguity
 - i.e., there is no vocabulary gap
- In short, an ingenious approach!
 - Seems full-proof (at least in theory)...

Easing the Burden of Search?

- controlling synonyms and near synonyms
 - one main heading/descriptor
 - leads to the preferred term
- control homographs
 - include qualification: bank (financial institute)
- provide scope notes
- displays broader, narrower and related terms
- express concepts elusive in free text

Index Languages

- Index language** is language used to describe documents and queries
- Exhaustivity**: number of different topics indexed
- Specificity**: level of accuracy of indexing
- Pre-coordinate indexing**: combinations of index terms (e.g. phrases) used as an indexing label
- Post-coordinate indexing**: combinations generated at search time

Controlling Vocabularies

- Three basic types
 - Subject Headings** is a list of preferred terms (descriptors)
 - precoordinated
 - including (near) synonyms (as non-descriptors)
 - Thesaurus** is an hierarchical arrangement (using BT/NT)
 - postcoordinated
 - may be faceted
 - Classification** is a systematic, hierarchical arrangement (coded)
 - precoordinated
 - coarse grained, usually no synonymy
- For IR, the specificity and # of terms assigned to documents seem to matter more than the precise system

Back to IR

- Long-standing debate on the indexing language
- Manual or human indexing:
 - indexers decide which keywords to assign to document based on controlled vocabulary
 - e.g. GOO, LCSH, MeSH, Yahoo!, Open Directories, . . .
 - significant cost (automatic assignment may be feasible)
- Automatic indexing:
 - indexing program decides which words, phrases or other features to use from text of document
 - indexing speeds are 10-100GB/hour for a single processor

Manual versus Automatic Indexing

- Original studies at Cranfield [Cleverdon, 1962, 1967]
 - automatic indexing can be at least as good as manual indexing
 - note: average performance, one-shot batch retrieval
- There is a stream of similar results
 - under average conditions free-text search outperforms controlled language search
- Perfect query gives perfect result
 - many queries (first shot) are not perfect
 - significantly hurts average performance

Choice of Search Terms is Crucial

- Searcher and indexer must pick the 'right' CV term
 - how to map an information need to the appropriate query?
 - this is a complex task for human indexers
 - can we expect end-users to do this effectively?
- Overlap between searchers [Iivonen, 1995]
 - overlap in chosen CV terms relatively low (only 30%)
 - overlap in 'concepts' much higher (up to 90%)
- Experienced searchers fare much better novices
 - intimate knowledge of CV is importance
 - pays off especially in interactive search

GIRT: Automatic Controlled Term Assignment

<title>Selbstbewusstsein von Mädchen <desc>Finde Dokumente, die über den Verlust des Selbstbewusstseins junger Mädchen während der Pubertät berichten.

<title>Self-confidence of girls <desc>Find documents which report on the loss of self-confidence of young girls during the puberty.

Ten closest used

Selbstbewußtsein
familiäre Sozialisation
Junge
Adoleszenz
Subkultur
Erziehungsstil
soziale Isolation
Marginalität
Bewußtseinsbildung
Pubertät

Ten closest global

Erwartung
Selbstbewußtsein
familiäre Sozialisation
Identitätsbildung
Identifikation
Sozialisationsbedingung
Junge
Adoleszenz
Freundschaft
Verhaltensmuster

Ten densest used

Erziehungsstil
Pubertät
Menstruation
körperliche Entwicklung
Selbsterstörung
Selbstbewußtsein
Heimerziehung
geistige Behinderung
Griechen
Bewußtseinsbildung

Results on Controlled Language Searching

- Expert searchers prefer CV [Hersh et al., 2001]
 - * Gives them control (structured queries)
 - * e.g., person as topic, person as author
 - * restricted search
 - * intuitive refinement of queries
 - * Usually low fall-out (few false negatives), may increase recall (think of patent search)
 - * Experts search with CV almost as effective as with free text!
- Free-text and CV search is often complementary
 - * retrieve different sets of relevant documents
 - * may help out if free text search fails (interactively)
 - * can we combine both approaches automatically?

Open Questions

- Open questions of [Svenonius, 1986]
 - * relative impact of free-text and controlled terms on retrieval?
 - * relative impact of different forms of control?
 - * if control is needed: what is its proper locus?
 - * IR system? Vocabulary? User?
 - * what are the relative costs?
- From scattered evidence on modern IR techniques:
 - * Free-text seems a viable alternative for most end-users
 - * In fact, ignoring free-text significantly hurts performance...

Unique Selling Points Remain

- Design can be tailor-made to needs
 - * customize vocabulary
 - * customize indexing depth
- Classifications (thesauri) provide overview and control
 - * signal activity/interest
 - * identify strengths/weakness by the 'paper' trails
- Uniform access to heterogeneous documents
 - * perfect for multilingual environments
 - * perfect for multimedia (non-text objects)
 - * easy for browsing (serendipity)

Amaryllis: Automatic versus Manual Assignment

<title>Impact sur l'environnement des moteurs diesel <desc>Pollution de l'air par des gaz d'échappement des moteurs diesel et méthodes de lutte antipollution. Emissions polluantes (NOX, SO2, CO, CO2, imbrûlés, ...) et méthodes de lutte antipollution

<title>The impact of diesel engine on environment <desc>Air pollution by the exhaust of gas from diesel engines and methods of controlling air pollution. Pollutant emissions (NOX, SO2, CO, CO2, unburned product, ...) and air pollution control

Manually assigned

Concentration et toxicité des polluants
Mécanisme de formation des polluants
Réduction de la pollution
Choix du carburant
Réglage de la combustion
Traitement des gaz d'échappement
Législation et réglementation

Automatically assigned

Moteur diesel
Qualité air
Azote oxyde
Norme ISO
Produit pétrolier
Lutte antipollution air
Véhicule à moteur
Gas oil
Consommation carburant
Carburant

Combining Free text and Controlled terms

- This is a battle-field in IR:
 - * In the early days various positive results
 - * E.g., expanding queries with controlled terms
 - * Context: little free text and poor IR models
 - * In modern IR combination rarely leads to significant improvement over free text
 - * Easy to disturb delicate balance of highly optimised models
- Some positive exceptions:
 - * Template-topics in TREC Genomics
 - * Complex feedback methods [Kamps, 2004]

Cyril Cleverdon 1967

- Cleverdon [1967, p.191–192] writes:

In practical all circumstances it would seem that such an [natural language] index language is more economical than any other; it will be interesting to see whether in some cases it also turns out to be more efficient.
- The cost/benefits argument is still valid
 - * Think of controlled vocabulary maintenance
 - * who's responsible? what if CV is changed? reassignment?

Wrap Up

- Subject Access to Digital Documents
 - * use free text (if available)!
 - * combine with controlled vocabulary (if available?)
- Controlled vocabularies
 - * range of unique advantages
 - * search requires more effort and knowledge in query formulation
 - * may pay-off for highly complex search requests, high recall tasks
 - * still appreciated by experts, minimal impact on naive end-users
- Lessons
 - * cost argument may be crucial
 - * trend toward "lightweight" subject access systems?
 - * only high-level, coarse grained?
 - * post-controlled vocabularies of [Lancaster, 1986]?

Recommended Reading

- A. Broder. A taxonomy of web search. *SIGIR Forum*, 36(2):3–10, 2002.
- C. W. Cleverdon. Report on the testing and analysis of an investigation into the comparative efficiency of indexing systems. Technical report, College of Aeronautics, Cranfield UK, 1962.
- C. W. Cleverdon. The Cranfield tests on index language devices. *Aslib*, 19:173–192, 1967.
- W. Hersh, A. Turpin, S. Price, D. Kraemer, D. Olson, B. Chan, and L. Sacherek. Challenging conventional assumptions of automated retrieval with real users: Boolean searching and batch retrieval evaluations. *Information Processing and Management*, 37:383–402, 2001.
- M. Iivonen. Consistency in the selection of search concepts and search terms. *Information Processing and Management*, 31:173–190, 1995.
- J. Kamps. Improving retrieval effectiveness by reranking documents based on controlled vocabulary. In S. McDonald and J. Tait, editors, *Advances in Information Retrieval: 26th European Conference on IR Research (ECIR 2004)*, volume 2997 of *Lecture Notes in Computer Science*, pages 283–295. Springer-Verlag, Heidelberg, 2004.
- F. W. Lancaster. *Vocabulary control for information retrieval*. Information Resources Press, Arlington VA, second edition, 1986.
- E. Svenonius. Unanswered questions in the design of controlled vocabularies. *Journals of the American Society for Information Science*, 37:331–340, 1986.