Indexing challenges in work place information retrieval

Controlled, human indexing vs full-text indexing

Are thesauri better used for query expansion than for controlled indexing?

Marianne Lykke Nielsen & Anna Gjerluf Eslau

NKOS 2006

Purpose of research project

- Focus of research project is use and performance of thesaurus in workplace information retrieval

- Evaluation and comparison of thesaurus as tool for
  - Information retrieval based on controlled, human indexing
  - Information retrieval based on full-text indexing, with thesaurus-based automatic query expansion

Case study

- Domain: pharmaceutical company H. Lundbeck (5000 employees)
- Retrieval system: Corporate document management system containing research documentation (25,384 items)
- Human indexing by use of facetted indexing policy and domain-specific thesaurus
- Thesaurus contains 5.200 concepts and 14.600 terms
- Searching by controlled metadata and full-text
- Clear and well-structured information needs
- Recall more important than precision

Methodology

How do indexing methods perform in retrieval?

- 10 real-life search cases
- Comparison of three search strategies:
    - Thesaurus-controlled metadata
    - Full-text
    - Full-text with QE
- Calculation of recall and precision
- Calculation based on relevance assessments by original searcher
- Scaled relevance assessments

Why did human indexing fail in retrieval?

- Analysis of documents assessed relevant for the 10 search jobs
- Analysis of internal, corporate indexing of search facets
- Identification and categorization of indexing problems causing low retrieval performance

Findings – performance

	Recall (%)										
Search strategy	SJ1	SJ2	SJ3	SJ4	SJ5	SJ6	SJ7	SJ8	SJ9	SJ10	Mean
Full-text	42	52	88	38	79	54	39	3	12	7	41
Full-text with QE (syn)	64	68	100	76	89	100	39	100	100	68	80
Full-text with QE (syn, nt)	100	90	100	87	89	100	39	100	100	73	88
Metadata	0	0	0	33	29	61	100	1	0	45	27

Findings – human indexing problems

Indexing problems	Frequency (%) N = 156	Explanations	
1. Conceptual analysis			
A1 Omission of topic	69	• Indexers fail to remember facets and topics that are not explicitly mentioned in indexing policy or checklist • Indexing policy recommend to check specific document sections such as title, table of content, etc. why indexers, especially in long documents, tend to omit topics from other document sections	√ ÷
A2 Misinterpretation and wrong perspective of topic	14	• Indexers misunderstand topic due to lack of topical and domain knowledge	÷
A3 Omission of implicit topic	2	• Difficult for indexers to determine degree of topical interpretation and domain-orientation	÷
2. Translation			
B1 Topic indexed at BT level	7		√
B2 Topic indexed with incorrect keyword	8	• Indexers misunderstand meaning and use of keywords	÷

Conclusions of case study

- Difficult to obtain complete, accurate and exhaustive human indexing
- Findings suggest that searching for specific topics should be based on full-text indexing, supported by thesaurus based query expansion
- Human indexing should focus on few, important, well-defined topics, e.g. used to develop taxonomies for broad browsing
- Analysis of relevance assessments indicates that full-text searches (with QE) might be improved by ranking, e.g. by
  - document type
  - publication year
  - Source
  - research approach