Indexing challenges in work place information retrieval

Marianne Lykke Nielsen and Anna Gjerluf Eslau

Royal School of Library and Information Science, Information Studies
Aalborg Branch, Sohngaardsholmsvej 2, 9000 Aalborg, Denmark
mln@db.dk

H. Lundbeck, Department of Regulatory Affairs
Ottliavej 9, 2500 Valby, Denmark
age@lundbeck.com

Introduction and research question

Information retrieval in workplace information environments is embedded in specific work task frameworks (Freund, Toms and Waterhouse, 2005). Contextual factors determine utility and relevance of retrieved documents. A domain-centred approach to indexing and retrieval seems necessary to meet the needs of the work domain (Mai, 2005).

Workplace environments are characterised by fluid collection definition, diversity in document types, and the need to search information from diverse repositories and file systems. Workplace retrieval systems cover a large variety of document formats, document genres, and languages, such as English and Danish and professional jargons (Abrol et al., 2001). Most workplace search technologies offer advanced indexing algorithms and search facilities, and there is an attempt to automate as much as possible, including indexing (Mahon, Hourican, Gilchrist, 2001). When indexing is carried out manually, it is often the document producers that perform the indexing task rather than professional indexers such as information specialists and librarians. Altogether, these characteristics provide another framework for indexing and information retrieval compared to traditional bibliographical retrieval systems in library settings from where we have most knowledge about the indexing process.

Several of the specific characteristics of workplace environments seem to be best met by controlled manual indexing compared to automatic indexing. The variety of document types challenge the use of automatic full-text indexing, as the search engine can not necessarily crawl and index the whole variety of document formats (Hawkings, 2004). The same applies to scanned documents, as the scanning may fail to read documents accurately. Diversity in language use is challenging when information retrieval is wanted across languages and jargons. An important disadvantage of automatic indexing is the lack of semantic control, which is an advantage of controlled indexing. The need to relate documents to specific work task frameworks may best be met by manual indexing as well, because the human indexer can judge the importance of the content and relate it to the context and situation. However, research comparing the qualities of the two basic approaches to

indexing: human, intellectual indexing and automatic, computer-based indexing, fails to provide conclusive results. It is difficult to determine whether performance differences are due to automatic versus human indexing, or to other variables such as the searching environment, the searcher's subject knowledge and searching expertise, the retrieval facilities, or the nature of the search. Depending on these variables, the indexing approaches produce lower or greater recall and precision (Rowley, 1994; Anderson & Pérez-Carballo, 2001ab; Savoy, 2004). There is a general recognition that the two should be used in combination to obtain the best search performance.

Research questions

The purpose of the present study was to expand upon previous investigations about indexing methods. The study was a case study within the context of workplace information retrieval. The project evaluated and explored how the two indexing methods, respectively controlled, manual indexing and computerised, automatic indexing perform in the context of work place retrieval systems, whether they are complementary as previous research concludes, and whether manual indexing meets better the specific characteristics and needs of workplace environment. The study also investigated the performance of query expansion, whether query expansion with use of a thesaurus improves automatic indexing, as preliminary research suggests (Rowley, 1994; Shiri & Revie, 2006).

The investigation was carried out as in two parts. First we carried out a retrieval test that focused on search effectiveness as measured by precision and recall. This part of the study was reported in Nielsen & Eslau (2006). The first test confirmed previous findings that automatic indexing improves recall. More surprising the investigation showed that extended query expansion with synonyms and narrower terms retrieved the largest number of highly relevant documents, including documents not retrieved by human indexing and simple full-text searching. The second part of the study that is reported in the present proposal investigated why the human indexing did not retrieve the highly relevant documents. The qualitative analysis seeked to identify the reason why the indexing failed, and in the study the indexing problems were defined and categorised. The study was explorative. The purpose was to provide an understanding of the indexing process and possible problems that may cause low retrieval performance.

Methodology

The overall purpose of the test design was to establish an evaluation framework that represents characteristics of real-life workplace information retrieval. The study was planned according to the framework of the tasks and situations of the case domain. Workplace documents were used as the document corpus, information needs that have been expressed to workplace search systems in real-life were the basis for development of search jobs, and the existing domain specific thesaurus was used for both controlled indexing with metadata and query expansion. Corporate end-users assessed the relevance of the retrieved documents from the perspective of the work domain, the basis for the calculation of precision and recall. Similarly, the qualitative analysis investigated the indexing failures from the perspective and needs of the case domain.

Findings

The indexing failures can be divided into five categories. Two categories, 'Authority requirements' and 'Perspective' are related to contextual factors; whereas general indexing issues can explain the three categories, 'Subject analysis', 'Translation' and 'Implicit subjects'. The paper discusses the five categories, especially how to counter the indexing problems. Some failures can be solved by the use of automatic indexing, whereas some indexing problems may still require human indexing.

Furthermore, the study showed that some descriptive metadata should be used only for description and sorting, because contextual metadata that is embedded in specific information tasks might leave out highly relevant documents from the hit list.

References

Abrol, M et al. (2001). Navigating large-scale semi-structured data in business portals. *Proceedings of the 27th VLDB Conference,* Roma, Italy, 2001.

Anderson, J D & Pérez-Carballo, J (2001a). The nature of indexing: how humans and machines analyze messages and texts for retrieval. Part I: Research, and the nature of human indexing. *Information Processing & Management*, 37. 231–254.

Anderson, J D & Pérez-Carballo, J (2001b). The nature of indexing: how humans and machines analyze messages and texts for retrieval. Part II. Maschine indexing, and the allocation of human versus maschine effort. *Information Processing & Management*, 37. 255–277.

Freund, L; Toms, E G & Waterhouse, J (2005). Modelling the information behaviour of software engineers using a work – task framework. Grove, A (eds.). *Proceedings 68th Annual Meeting of the American Society for Information Science and Technology*, Charlotte, NC, October 28 – November 3, 2005.

Hawking, D (2004). Challenges in enterprise search. *Proceedings of the Australian Database Conference*, Dunedin, New Zealand, 2004.

Lancaster, W (2003). *Indexing and abstracting in theory and practice*. London: Facet publishing.

Mahon, B, Hourican, R & Gilchrist, A (2001). *Research in information architecture.* London: TFPL.

Mai, J-E (2005). Analysis in indexing: document and domain centred approaches. *Information Processing and Mangement,* 41. 599–611.

Nielsen, M L & Eslau, A G (2006). Human versus automatic indexing in the context of workplace retrieval systems. (Submitted for review).

Rowley, J (1994). The controlled versus natural indexing languages debate revisited: a perspective on information retrieval practice and research. *Journal of Information Science*. 20(2). 108–119.

Savoy, J (2004). Bibliographic database access using free-text and controlled vocabulary: an evaluation. *Information Processing & Management*, 41. 873–890.

Shiri, A & Revie, C (2006). Query expansion behaviour within a thesaurus-enhanced search environment: a user-centered evaluation. *Journal of the American Society for Information Science*. 57(4). 462–478.