Testing the principle of polyrepresentation

Mette Skov, Henriette Pedersen, Birger Larsen and Peter Ingwersen
Department of Information Studies, Royal School of Library and Information Science
Birketinget 6, DK-2300 Copenhagen S, Denmark
Email: {ms,blar,pi}@db.dk

Categories and subject descriptors: H.3.3. Information Search and Retrieval; H.3.7. Digital Libraries

General Terms: Performance; Experimentation.

Keywords: Information Retrieval, Polyrepresentation; Cognitive Overlaps.

1. INTRODUCTION

The cognitive theory of Information Retrieval (IR) and the principle of polyrepresentation derived from it [1, 2] provide a theoretical background for how to exploit different document features in order to improve performance in IR. In summary, the theory hypothesises that overlaps between different cognitive representations of both users' information situation as well as documents can be exploited for reducing the uncertainties inherent in IR, thereby improving the performance of IR systems. Good results are expected when cognitively unlike representations are used, e.g., the document title (made by the author) vs. intellectually assigned descriptors from indexers and vs. citations made by other authors over time.

Essentially the principle of polyrepresentation signifies to make use of a variety of *contexts* – in particular associated with information objects – but also in principle related to the searcher and other central components of interactive IR. Context in IR can take several forms [6]. Intra-document structures and representations (signs) are contextual to one another, as are the documents themselves (inter-document relationships). The current session constitutes a third kind of context dealing with features of the interaction between searcher(s) and documents. Examples are eye or mouse movements and pointing and searcher request features, like depth of knowledge on work task. Further, the interaction processes are in context of the conceptual, emotional, systemic and socio-organizational properties immediately surrounding the searcher and documents. All actors, information systems, documents and interactive sessions are influenced to a certain extent by remote contextual constructs, such as general techno-economic and socio-cultural infrastructures in society. Across this stratification operates an additional dimension, that of the historic context of actors' experiences forming their expectations. In the present paper elements of the intra and inter-document contexts are involved, that is, functionally different content representations like titles and abstracts and cognitively different representations like descriptors assigned by indexers as well as inter-document relationships in the form of bibliographic references made by the author.

When cognitively (and functionally) different representations point at the same documents via so-called cognitive overlaps, it is regarded as evidence of high probability of relevance. In this paper we present the results from an experiment testing the principle of polyrepresentation in a test collection. The main purpose of the experiment is to show whether the use of cognitively different document representations as suggested in the cognitive theory of IR can enhance performance. Although the cognitive theory of IR and the principle of polyrepresentation by nature are holistic, polyrepresentation is, however, inherently Boolean in much of its reasoning. This is apparent in the pronounced focus on cognitive retrieval overlaps, i.e., sets of retrieved documents based on different cognitive representations. The second purpose of this experiment is to show the consequences of implementing a highly structured search strategy into a best match IR system.

2. THE EXPERIMENT

The test environment was the Cystic Fibrosis test collection [5] indexed in the InQuery retrieval system (version 3.1). The test collection consists of 1,239 document representations from MedLine, 100 requests and tripartite relevance assessments. This test collection is ideally suited for testing polyrepresentation as it contains two cognitively different representations and a number of functionally different ones. Hence, we made use of the functionally different titles (TI), abstracts (AB) and references (RF) all made by the author. Major (MJ) and minor MeSH (MN) descriptors used in the experiment are functionally different, both representing the indexer as a cognitive agent. In order to test the principle of polyrepresentation we made use of 29 requests searched as both natural language and highly structured queries. The former queries were weakly structured [3] in the sense that Boolean operators were used only to combine the representations for identifying overlaps in InQuery. The highly structured queries were structured by the use of InQuery's Boolean operators in combining query facets and also in combining representations for identifying overlaps. Furthermore proximity operators were used to search phrases and MeSH synonyms were added to the search keys. By using both natural language and highly structured queries it was possible to analyse which search configuration was optimal when implementing polyrepresentation in a best match system.

The experiment revealed that on average 88% of documents found searching AB were also found searching TI. Therefore, the two representations (TI/AB) were indexed in the same field. Combining the four representations (TI/AB, MJ, MN and RF) resulted in 15 overlaps[1] (see table 1). The overlaps were defined such that a document could appear in one, and only one, of the 15 overlaps. Inspired by an earlier study of polyrepresentation [4] references were included in the search without a priory intellectual selection of

SIGIR'04, July 26-29, 2004, Sheffield, UK.
ACM 1-58113-000-0/00/0004…$5.00

[1] We use the term overlap even though the documents in overlaps 12-15 were retrieved form one representation only.

seed documents. Instead a subject search was performed for each request in SciSearch. The cited references in the retrieved documents were ranked using the RANK command in Dialog. For each request the cited references ranked top three on the list were used as input in a (RF) search in the test collection. Those search results for the reference representation included documents containing one or more of the top three cited references in their reference lists. The 29 requests from the test collection were used without modification as direct bag-of-words input for the natural language queries searched in TI/AB, MJ and MN, respectively. The highly structured queries consisted of the same 29 requests, modified in a number of ways. First, a noun-phrase finder parsed them. Secondly, stop words were removed, and finally the remaining search keys were expanded intellectually using the MeSH thesaurus.

3. RESULTS AND DISCUSSION

The tripartite relevance assessments provided in the test collection made it possible to investigate the retrieval performance for both all relevant documents (relevant and highly relevant documents) and highly relevant documents. The results are presented in table 1. Columns A and D show the number of documents found in each of the 15 overlaps in natural and highly structured language. Not surprisingly the natural language queries result in many more documents than do highly structured queries. Table 1 also shows that, in general, overlaps generated from three or four representations (overlap 1-5) have higher precision than overlaps generated from two or one representations (overlap 6-11 over overlap 12-15). These findings support the principle of polyrepresentation suggesting that a high number of representations pointing towards a document are likely to be an indicator of high precision. Table 1 shows that for all 15 overlaps highly structure queries result in higher precision (column E) than queries in natural language (column B). From a polyrepresentative point of view this can be explained by looking at the query structure. The highly structured queries ensure that documents identified in an overlap have identical or synonym search terms present from all the representations searched. The weak structure in the natural language queries does not ensure that the search terms (or synonyms) are present in each of the document sets generating the overlaps. Therefore, proper polyrepresentation in the true sense of the concept cannot be achieved with weakly structured queries in natural language.

As described above, table 1 reveals, that overlaps generated from three or four representations have higher precision than overlaps generated from two representations etc. However, looking at the overlaps generated from 3 representations (overlap 2-5) in the highly structured queries, overlap 2 has a considerably lower precision than overlaps 3-5. This suggests that the RF as representation is important to obtain high precision. This high-precision trend when RF is included also pertains to overlaps generated from 2 different representations (overlaps 8, 10, 11). These findings stress the importance of including representations that are both cognitively dissimilar (TI/AB; MJ/MN) and functionally different (RF) as suggested in the cognitive theory of IR.

4. CONCLUSION

The experiment supports the principle of polyrepresentation that states that when a number of cognitively (and functionally) different representations point to a document, this fact is likely to be an indicator of high relevance. Because of the structured (Boolean) nature of polyrepresentation, a strongly structured query language is necessary when implementing the principle of polyrepresentation in a best match IR system. Finally, the results indicate that scientific references (outlinks) serve as central contextual elements during IR and constitute an important type of representation in order to obtain high precision.

Table 1: Recall and precision for the 15 overlaps. (Ol = overlap, P = precision, R = recall, # doc. = number of retrieved documents).

	Natural language			Highly structured language				
	# doc.	P all relevant	R all relevant	# doc.	P all relevant	P highly relevant	R all relevant	R highly relevant
Overlap	A	B	C	D	E	F	G	H
Ol 1 (ti/ab, mj, mn, rf)	126	41%	5%	58	69%	53%	4%	6%
Ol 2 (ti/ab, mj, mn)	668	13%	8%	100	42%	20%	4%	4%
Ol 3 (ti/ab, mj, rf)	101	48%	4%	66	79%	45%	5%	6%
Ol 4 (ti/ab, mn, rf)	240	29%	6%	68	62%	47%	4%	7%
Ol 5 (mj, mn, rf)	3	0	0	11	64%	45%	1%	1%
Ol 6 (ti/ab, mj)	702	12%	7%	131	45%	22%	5%	6%
Ol 7 (ti/ab, mn)	1761	9%	14%	210	27%	13%	5%	6%
Ol 8 (ti/ab, rf)	1528	9%	12%	162	27%	19%	4%	6%
Ol 9 (mj, mn)	141	6%	1%	42	26%	14%	1%	1%
Ol 10 (mj, rf)	6	33%	0	16	38%	19%	1%	1%
Ol 11 (mn, rf)	42	21%	1%	68	34%	16%	2%	2%
Ol 12 (ti/ab)	16201	2%	25%	770	12%	5%	8%	8%
Ol 13 (mj)	106	10%	1%	109	27%	12%	3%	3%
Ol 14 (mn)	603	4%	2%	336	17%	7%	5%	5%
Ol 15 (rf)	872	5%	0	2458	6%	2%	12%	10%

5. ACKNOWLEDGMENTS

The authors wish to thank the Center for Intelligent Information Retrieval, University of Massachusetts Computer Science Department, Amherst, MA for providing the InQuery software

6. REFERENCES

[1] Ingwersen, P. (1994): Polyrepresentation of information needs and semantic entities: elements of a cognitive theory for information retrieval interaction. In: *Proceedings of SIGIR 1994,* p. 101-110.

[2] Ingwersen, P. (1996): Cognitive perspectives of information-retrieval interaction - elements of a cognitive IR theory. *Journal of Documentation*, 52(1), p. 3-50.

[3] Kekäläinen, J. and Järvelin, K. (1998): The impact of query structure and query expansion on retrieval performance. In: *Proceedings of SIGIR 1998,* p. 130-137.

[4] Larsen, B. and Ingwersen, P. (2002): The boomerang effect: retrieving scientific documents via the network of references and citations. In: *Proceedings of SIGIR 2002,* p. 397-398. (poster paper)

[5] Shaw, W. M.; Wood, J. B., Wood, R. E., and Tibbo, T. R. (1991). The cystic fibrosis database: content and research opportunities. *Library and Information Science Research,* 13, p. 347-366. (The test collection may be downloaded from http://www.dcc.ufmg.br/irbook/cfc.html).

[6] Ingwersen, P. and Järvelin, K. (Forthcoming). *Integration of Information Seeking and Retrieval in Context.* Kluver.