In: Ruthven, I., Borlund, P., Ingwersen, P., Belkin, N., Tombros, A. & Vakkari, P. (eds.), *Information Interaction in Context*: Proceedings of the International Symposium on Information Interaction in Context (IiiX 2006). Copenhagen, New York: Royal School of Library and Information Science /ACM Press, 2006: 163-170.

Inter and Intra-Document Contexts Applied in Polyrepresentation

Mette Skov, Birger Larsen and Peter Ingwersen

Department of Information Studies, Royal School of Library and Information Science
Birketinget 6, DK–2300 Copenhagen S, Denmark
{ms, blar, pi}@db.dk

Abstract. The principle of polyrepresentation offers a theoretical framework for handling multiple contexts in Information Retrieval (IR). This paper presents an empirical study of polyrepresentation of the information space with focus on inter and intra-document features. The Cystic Fibrosis test collection indexed in a best match system constitutes the experimental setting. Overlaps between four functionally and/or cognitively different representations are identified. Supporting the principle of polyrepresentation, results show that in general overlaps generated by three or four representations have higher precision than those generated from one or two overlaps both in structured and unstructured search mode. It is concluded that a highly structured query language is necessary when implementing the principle of polyrepresentation in a best match IR system because the principle is inherently Boolean. Finally a re-ranking test shows promising results when search results are re-ranked according to precision obtained in the overlaps.

Symposium themes. Document structure in contextual IIR

1. Introduction

Based on the cognitive approach to IR *the principle of polyrepresentation* is introduced by Ingwersen [1] and further elaborated in Ingwersen & Järvelin [2]. Essentially the principle of polyrepresentation signifies to make use of a variety of *contexts* associated with the interactive IR process - in particular contexts associated with the information object but also contexts related to the searcher.

The principle of polyrepresentation is based on the following main hypothesis: "…the more interpretations of different cognitive and functional nature, based on an IS&R [Information Seeking & Retrieval] situation, that point to a set of objects in so-called cognitive overlaps, and the more intensely they do so, the higher the probability that such objects are *relevant* (pertinent, useful) to a perceived work task/interest to be solved, the information (need) situation at hand, the topic required, or/and the influencing context of that situation…" [2, p. 208]. The interpretations take form of different representations of context e.g. the document title, intellectually assigned descriptors from indexers and citations (in-links).

A small number of empirical studies have so far been based on the principle of polyrepresentation. The studies involve in very different ways a variety of contexts - illustrating the holistic nature of polyrepresentation. Kelly et al. [5] investigated polyrepresentation of the user's cognitive space by combining different searcher statements of a single information need. Lund [7] examined the retrieval results from the 12 most effective TREC 5 search engines. In Lund's study the search engines illustrate different representations of IR system settings. A third example of an empirical study of polyrepresentation was conducted by Larsen [6]. Larsen's study is an example of polyrepresentation of information space and involved different inter and intra-document representations from the INEX test collection and INSPEC thesaurus. The so-called Boomerang Effect was tested by applying citation cycling strategies, i.e., backward chaining followed by forward citation chaining. Finally, Skov et al. [9] tested the principle of polyrepresentation in a best match setting and like Larsen [6] the focus was on elements of inter and intra-document contexts. Intra-document contexts were involved in the form of functionally different content representations like titles and abstracts and cognitively different representations like descriptors assigned by indexers. Inter-document relationships were involved in the form of bibliographic references and citations.

This paper elaborates on the results of Skov et al. [9] and the main purpose of the paper is to show whether the use of cognitively different representations as suggested in the principle of polyrepresentation can enhance retrieval performance. Secondly, we want to test if re-ranking of retrieved documents according to obtained precision in overlaps between different representations *or* re-ranking based on citation impact can improve performance.

The paper is structured as follows: Section 2 outlines the experimental setting and places the four document representations included in the experiment in a polyrepresentation framework. Section 3 presents and discusses the results of the experiment. Section 4 suggests future possible research on polyrepresentation.

2. Testing inter- and intra-document features in polyrepresentation

Skov et al. [9] tested elements of polyrepresentation of the information space in a best match setting. The Cystic Fibrosis test collection [8] indexed in the probabilistic InQuery[1] retrieval system (version 3.1) constituted the experimental setting. The Cystic Fibrosis test collection contains 1,239 documents from Medline[2] and graded relevance assessments (retrieved documents were judged highly relevant, marginally relevant or not relevant). Table 1 shows summary figures for the Cystic Fibrosis test collection.

[1] The InQuery software was provided by the Center for Intelligent Information Retrieval, University of Massachusetts Computer Science Department, Amherst, MA, USA.

[2] Medline is the US National Library of Medicine's bibliographic database (http://www.ncbi.nlm.nih.gov/entrez/query.fcgi).

Table 1. Summary figures for the Cystic Fibrosis test collection

Summary figures	All (highly + marginally) relevant documents	Highly relevant documents
Number of topics	29	29
Total number of relevant documents in the collection	1134	490
Minimum number of relevant documents per topic	6	3
Maximum number of relevant documents per topic	177	93
Median	30	12
Mean	39,10	16,90
Standard deviation	33,21	17,34

The Cystic Fibrosis test collection is small, however, it is ideally suited for testing inter and intra-document features in polyrepresentation as it contains both cognitively and functionally different features. Skov et al. [9] applied functional different features in the form of titles (TI), abstracts (AB) and references (RF) all representing the author combined with Medical Subject Headings (MeSH) representing the indexer as a cognitive agent. MeSH is a controlled vocabulary used for indexing, and MJ and MN represent the major and minor subjects of the document respectively. The representations TI and AB were merged and searched as one representation (TI/AB). Combining the four representations (TI/AB, MJ, MN and RF) resulted in 15 overlaps[3] (table 2). Inspired by an earlier study of polyrepresentation [6] references were included in the search without *a priory* intellectual selection of seed documents. Instead a subject search was performed for each request in SciSearch. The references in the retrieved documents were ranked using the RANK command in Dialog. For each request the top three cited references were used as input in a (RF) search in the test collection.

Skov et al. used 29 topics and two types of queries were tested: unstructured, natural language queries and highly structured queries. The 29 requests were used without modification as direct bag-of-words input for the natural language queries searched in TI/AB MJ and MN, respectively, and combined with Boolean logic. The same 29 requests were searched as highly structured queries, modified in a number of ways inspired by Kekäläinen & Järvelin's approach to query structuring [4]: First, queries were parsed for noun-phrases. Secondly, stop words were removed, and finally the remaining search keys were expanded manually using the MeSH thesaurus. Inspired by Kekäläinen & Järvelin [4] we used boolean operators to express relations between search terms in highly structured queries. The following examples of a highly structured query are based on an example request: *What are the hepatic complications of cystic fibrosis?* First an example conceptual query where the terms in italic are expanded terms from MeSH. The example illustrates an overlap 2 search where overlap between TI/AB, MJ, MN is identified:

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

```
TI/AB = (hepatic OR liver OR hepatectomy OR "hepatic complications")

AND MJ = (hepatic OR liver OR hepatectomy OR "hepatic complications")

AND MN = (hepatic OR liver OR hepatectomy OR "hepatic complications")

NOT RF = (Bowman T Lancet 1 183 962, Weber A Pediatrics 4 53 949)
```

The same example query expressed in the query language of InQuery:

```
#q = #bandnot[4] (#band (#field (TI/AB #syn[5] (hepatic liver hepatectomy
#1[6]("hepatic complications))), #field (MJ #syn (hepatic liver
hepatectomy #1("hepatic complications))), #field (MN #syn (hepatic
liver hepatectomy #1("hepatic complications)))), (#field (RF #1
(Bowman T Lancet 1 183 962) #1 (Weber A Pediatrics 4 53 949)))
```

Using query structure as a test variable provides information on the impact of query structure when applying polyrepresentation in a best match setting.

3. Results and discussion

The tripartite relevance assessments provided in the test collection made it possible to investigate the retrieval performance for both (a) all relevant documents (relevant and highly relevant documents) and (b) highly relevant documents. The results show that, in general, overlaps generated by three or four representations have higher precision than those generated from one or two overlaps both in structured and unstructured search mode (table 2). 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 [1]. For all 15 overlaps highly structured queries result in higher precision than queries in natural language supporting the Kekäläinen & Järvelin [4] findings (Table 2). From a polyrepresentative point of view this can be explained by looking at the two different query structures. The highly structured queries tend to ensure that documents identified in an overlap have identical or synonym search terms present from all the representations searched. In contrast the natural language queries only require one search term from the query to be present, and as a consequence the retrieved sets and overlaps between them include document representations with no or only little relation to the information need. Therefore, polyrepresentation in the true sense of the concept cannot be achieved with weakly structured queries in natural language.

Differences are found between representations; in particular the results in table 2 indicate an increase in precision when documents retrieved from the reference (RF) form part of an overlap. On the other hand precision drops when MN terms contribute

[4] Bandnot and band in InQuery query language are equivalent to Boolean NOT and AND respectively.

[5] The keys within a #syn operator are treated as instances of the same term.

[6] The keys within a #1 operator must be found within one word of each other.

to an overlap (overlap 7, 11, and 14). Both findings stress the importance of using representations that are both cognitively dissimilar (e.g., TI/AB and MeSH-headings) and functionally different (e.g., references).

Table 2. Recall and precision for the 15 overlaps. Table 2 is based on Skov et al. [9]. Ol = overlap, P = precision, R = recall, # doc. = number of retrieved documents. Recall = 0 signifies values below 0,005%

	Natural language queries			Highly structured queries				
Overlap	# doc.	P all relevant	R all relevant	# doc	P all relevant	P highly relevant	R all relevant	R highly relevant
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%	4%	2458	6%	2%	12%	10%

In large scale databases ranking of highly relevant documents in top of the search result is important to end users. In two different re-ranking tests elements from the principle of polyrepresentation was applied as a way to increase performance measured by Cumulative Gain (CG) [3]. The first re-ranking test (run 2-4) was based on obtained precision in highly structured queries in the 15 overlaps. First a fusion of search results from the 15 overlaps into one merged search result was carried out. Next the overlaps were weighted according to (a) precision obtained in the highly structured queries and (b) precision obtained retrieving highly relevant documents. The following weights were applied: In run 1 no weights were applied; in run 2 weight 100 was applied to the four overlaps with the highest precision (overlap 1, 3, 4, and 5); in run 3 weight 100 was applied to the four overlaps with highest precision and weight 50 was applied to overlaps with medium precision (overlap 2, 6, 8, and 10); in run 4 weight 100 was applied to overlap 1 containing four different representations, weight 66 was applied to overlap 2, 3, 4, and 5 containing three different representations, weight 33 was applied to overlap 6, 7, 8, 9, 10, and 11

including two different representations. Table 3 shows CG (document cut-off value = 30).

Table 3. Cumulative Gain: run 1-6 and bag-of-words.

Rank	Ideal vector	Run 1	Run 2	Run 3	Run 4	Run 5	Run 6	Bag-of-words
5	9.8	4.5	5.5	5.3	4.9	5.3	5.4	5.9
10	18.1	8.4	10.2	9.7	8.9	9.6	9.5	10.1
15	24.8	11.2	13.6	13.4	12.0	12.9	12.6	13.0
20	30.7	13.6	16.8	15.5	14.3	15.8	15.6	14.9
25	35.1	15.2	18.5	18.2	16.7	17.7	17.0	16.9
30	38.7	17.1	20.2	20.0	18.2	19.3	18.7	18.6

Run 1: No weighting applied

Run 2: Overlap 1, 3, 4, and 5: weight 100.

Run 3: Overlap 1, 3, 4, and 5: weight 100; overlap 2, 6, 8, and 10: weight 50

Run 4: Overlap 1: weight 100; overlap 2, 3, 4, and 5: weight 66;
overlap 6, 7, 8, 9, 10, and 11 weight 33

Run 5: Overlap 1, 3, 4, and 5: weight 100 + received at least one citation

Run 6: Overlap 1, 3, 4, and 5: weight 100 + received at least three citations

To decide whether differences in CG figures are statistically significant; we ran the Friedman test, which is based on ranks. The test was based on normalised average CG vectors as suggested by Järvelin and Kekäläinen [3]. The Friedman test showed that search performance for baseline bag-of-words and runs 2-4 (weights applied to high precision overlaps) was significantly better than run 1 where no weights are applied ($df = 4$, $p < 0.05$). Results in table 3 indicate that run 2 performs slightly better than run 3-4, however, the differences between run 2-4 are not statistical significant (probably partly caused by the small test collection). Comparing weighting according to polyrepresentation (run 2-4) and InQuery's weighting of natural language queries (bag-of-words) shows no statistical significant difference. This suggests that weighting according to the principle of polyrepresentation performs equally good as InQuery's weighting.

The second re-ranking test (run 5 and 6) was based on citation impact; applying citations as an indication of quality. Citations (in-links) represent the citing author's cognitive structures and their interpretations of the work cited. In run 5 and 6 all documents included in the search result had received at least one or three citations (respectively). Results in table 3 indicate a slight decrease in retrieval performance when applying citations as an indication of quality compared to the best run of weighted overlaps (run 2). However, the Friedman test showed no statistical significance between baseline bag-of-words and runs including citations (run 5 and

6). Table 3 also shows that InQuery's ranking (bag-of-words) in top 15 outperform run 1-6. However, looking at rank 15-30 both run 2, 5 and 6 outperform InQuery's weighting. The Cystic Fibrosis test collection does not include citation titles and therefore citations are not used according to the principle of polyrepresentation [1] in the present study. In addition, the volume of citations is small making re-ranking problematic.

4. Conclusions and future work

In general the experiment supports the principle of polyrepresentation suggesting that a high number of cognitively and functionally different representations pointing towards a document are likely to be an indicator of high precision. Because of the Boolean nature of polyrepresentation, a strongly structured query language is necessary when implementing the principle of polyrepresentation in a best match IR system. The results indicate that scientific references are an important type of representation in order to obtain high precision. Finally, re-ranking tests showed statistically significant improvements when weights were applied to high precision overlaps. Results from re-ranking test based on citation impact indicate a slight decrease in performance.

Future research could include work on combining the representations using semi-structured best match queries. Such an approach could retain some of the structure of the Boolean queries, while softening the rigidity of them. Future testing of the principle of polyrepresentation should apply larger scale evaluation. The test collection from the TREC web track would potentially be a possibility. This opens new ways of exploring polyrepresentation as anchor text and URL text can represent intra-document context and hyperlinks can represent inter-document context. Another possibility is the large collection of Medline records used by the TREC genomics track; however, no references or citations are included in this collection.

References

1. Ingwersen, P.: Cognitive perspectives of information-retrieval interaction - elements of a cognitive IR theory. Journal of Documentation, Vol. 52, 1 (1996) 3-50
2. Ingwersen, P. & Järvelin, K.: The Turn: Integration of Information Seeking and Retrieval in Context. Springer, Dordrecht (2005)
3. Järvelin, K. & Kekäläinen, J.: Cumulated Gain-based evaluation of IR techniques. ACM Transactions on Information Systems, Vol. 20, 4 (2002) 422-446.
4. Kekäläinen, J. & Järvelin, K.: The impact of query structure and query expansion on retrieval performance. In: Croft, W.B., Moffat, A., van Rijsbergen, C.J., Wilkinson, R. & Zobel, J. (eds.): Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM Press, Melbourne, Australia, New York, NY (1998) 130-137

5. Kelly, D., Dollu, V.D. & Xin Fu.: The loquacious user: A document-independent source of terms for query expansion. In: Proceedings of the 28th Annual ACM SIGIR Conference on Research and Development in Information retrieval. ACM Press, New York NY (2005) 457-464
6. Larsen, B.: References and Citations in Automatic Indexing and Retrieval Systems: Experiments with the Boomerang Effect. The Royal School of LIS, Copenhagen DK, Ph.D. Thesis (2004) [Available at: http://www.db.dk/blar/dissertation . Cited August 2, 2006.]
7. Lund, B.R.: Polyrepræsentation og Datafusion: Test af teorien om polyrepræsentation gennem forsøg med fusion af TREC-5 resultater [Fusion of TREC-5 results – testing the principle of polyrepresentation]. Royal School of LIS, Copenhagen, MSc Thesis (2005)
8. Shaw, W. M.; Wood, J. B., Wood, R. E., & Tibbo, T. R.: The cystic fibrosis database: content and research opportunities. Library and Information Science Research, 13 (1991) 347-366
9. Skov, M., Pedersen, H., Larsen, B. & Ingwersen, P.: Testing the principle of polyrepresentation. In: Larsen, B. (Ed.) Information Retrieval in Context, Proceedings of the SIGIR 2004 IRiX Workshop. Sheffield UK, July 2004: (2004) 47-49. [Available at http://ir.dcs.gla.ac.uk/context/IRinContext_WorkshopNotes_SIGIR2004.pdf . Cited August 2, 2006.]