Testing the Principle of Polyrepresentation

M. Skov, H. Pedersen, B. Larsen and P. Ingwersen
Department of Information Studies
Royal School of Library and Information Science
Denmark

Outline

- Polyrepresentation and context
- Context within and between documents
- An experiment on a small test collection
  - Aims and motivation
  - Data and methods
  - Results
- Conclusions and future work

Polyrepresentation and context

- IR in context = a *lot* of contexts
- Some of these must be transformed into tangible representations to do IR
- …but which ones to use, and in what combinations?
- The principle of polyrepresentation (Ingwersen, 1996) might serve as a framework because of its holistic nature

Polyrepresentation and context

- Polyrepresentation 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
- High precision is expected when (several) cognitively and functionally unlike representations point to the same documents
- For instance, the document title (made by the author) vs. intellectually assigned descriptors from indexers vs. citations made by other authors over time

An experiment

- Aim: to do an empirical test of elements of polyrepresentation by exploiting different contexts within and between documents
  - Can overlaps between representations enhance performance, in particular with respect to precision?
- Motivation: This hadn't really been done explicitly based on the principle of polyrepresentation before
- In addition, polyrepresentation is inherently Boolean
  - A secondary aim was to investigate the consequences of implementing the principle in a best match system

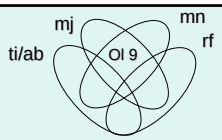
Data and methods

- The Cystic Fibrosis test collection (Shaw et al., 1991)
  - 1,239 records, 100 requests + tripartite relevance assessments
- Very small by today's standards,
- …but interesting because it contains representations derived from a number of contexts
- This diversity is essential for the experiment, but cannot be found in TREC or other collections

Data and methods

- Representations used: Title/Abstract, References, Major MeSH, and Minor MeSH
- Queries used: 29 of 100 in two versions
  - Natural language (bag-of-words)
  - Highly structured
    - Automatically identified phrases
    - Expansion with MeSH synonyms
    - Query facets, inspired by Kekäläinen and Järvelin (1998)
  - *Seed documents* were identified in Science Citation Index for both
- All 15 possible overlaps (TI/AB;RF;MJ;MN) were identified in both query types using InQuery

Results



		Natural language			Highly structured				
			All relevant			All relevant		Highly relevant	
		# doc	Precision	Recall	# doc	Precision	Recall	Precision	Recall
Overlap		A	B	C	D	E	G	F	H
Ol 1	(ti/ab,mj,mn,rf)	126	41%	5%	58	69%	4%	53%	6%
Ol 2	(ti/ab,mj,mn)	668	13%	8%	100	42%	4%	20%	4%
Ol 3	(ti/ab,mj,rf)	101	48%	4%	66	79%	5%	45%	6%
Ol 4	(ti/ab,mn,rf)	240	29%	6%	68	62%	4%	47%	7%
Ol 5	(mj,mn,rf)	3	0%	0%	11	64%	1%	45%	1%
Ol 6	(ti/ab,mj)	702	12%	7%	131	45%	5%	22%	6%
Ol 7	(ti/ab,mn)	1761	9%	14%	210	27%	5%	13%	6%
Ol 8	(ti/ab,rf)	1528	9%	12%	162	27%	4%	19%	6%
Ol 9	(mj,mn)	141	6%	1%	42	26%	1%	14%	1%
Ol 10	(mj,rf)	6	33%	0%	16	38%	1%	19%	1%
Ol 11	(mn,rf)	42	21%	1%	68	24%	2%	16%	2%
Ol 12	(ti/ab)	16201	2%	25%	770	12%	8%	5%	8%
Ol 13	(mj)	106	10%	1%	109	27%	3%	12%	3%
Ol 14	(mn)	603	4%	2%	336	17%	5%	7%	5%
Ol 15	(rf)	872	5%	0%	2458	6%	12%	2%	10%

Results

- Highly structured queries resulted in less documents than natural language queries, but they showed higher precision
- Overlaps generated from three or four representations showed higher precision than those generated from two or one (in particular for highly structured queries)

Results

			Natural language			Highly structured				
				All relevant			All relevant		Highly relevant	
			# doc	Precision	Recall	# doc	Precision	Recall	Precision	Recall
	Overlap		A	B	C	D	E	G	F	H
4	Ol	1 (ti/ab,mj,mn,rf)	126	41%	5%	58	69%	4%	53%	6%
	Ol	2 (ti/ab,mj,mn)	668	13%	8%	100	42%	4%	20%	4%
3	Ol	3 (ti/ab,mj,rf)	101	48%	4%	66	79%	5%	45%	6%
	Ol	4 (ti/ab,mn,rf)	240	29%	6%	68	62%	4%	47%	7%
	Ol	5 (mj,mn,rf)	3	0%	0%	11	64%	1%	45%	1%
	Ol	6 (ti/ab,mj)	702	12%	7%	131	45%	5%	22%	6%
	Ol	7 (ti/ab,mn)	1761	9%	14%	210	27%	5%	13%	6%
2	Ol	8 (ti/ab,rf)	1528	9%	12%	162	27%	4%	19%	6%
	Ol	9 (mj,mn)	141	6%	1%	42	26%	1%	14%	1%
	Ol	10 (mj,rf)	6	33%	0%	16	38%	1%	19%	1%
	Ol	11 (mn,rf)	42	21%	1%	68	24%	2%	16%	2%
	Ol	12 (ti/ab)	16201	2%	25%	770	12%	8%	5%	8%
1	Ol	13 (mj)	106	10%	1%	109	27%	3%	12%	3%
	Ol	14 (mn)	603	4%	2%	336	17%	5%	7%	5%
	Ol	15 (rf)	872	5%	0%	2458	6%	12%	2%	10%

Results

- Highly structured queries resulted in less documents than natural language queries, but had higher precision
- Overlaps generated from three or four representations showed higher precision than those generated from two or one (in particular for highly structured queries)
- For overlaps generated by three representations, those involving references display considerably higher precision

Conclusion

- Overall, the principle of polyrepresentation is supported
  - more representations → higher precision
- However, this was only consistent for the highly structured queries.
  - The weak structure in the natural language queries does not ensure that the search keys are present in of the sets → poor coordination and loss of precision

Future work

- Verify the results on a larger collection
  - INEX (with descriptors from INSPEC) is a candidate
- Test a larger range of representations
- Test different combinations of document representations with representations of the user's context
- Reduce the amount of manual/intellectual work
  - Use automatic methods more
  - Generate overlaps automatically for *n* sets

Thank you!