Using value-added document representations in INEX

Birger Larsen, Haakon Lund, Jacob K. Andresen and Peter Ingwersen

Department of Information Studies

Royal School of Library and Information Science

Birketinget 6, DK-2300 Copenhagen S, Denmark

{blar,hl,jaa,pi}@db.dk

ABSTRACT

Viewing Information Retrieval from the cognitive viewpoint we generated different functional and cognitive representations of the INEX corpus using both the XML structure and external sources. This included the use of a citation index and intellectually assigned descriptors, and an expansion of the document representation through a domain thesaurus. The aim was to investigate the possible benefits from applying the principle of polyrepresentation [3]. Results showed that neither the descriptors nor the expanded document representation through the thesaurus could improve results with the natural language queries used. The citation index achieved a similar performance to that obtained using various kinds of titles extracted from the XML structure.

1 INTRODUCTION

Highly structured XML documents offer unique opportunities for extracting many different representations of documents for Information Retrieval (IR) purposes. In this paper we describe our efforts to work with combinations of different representations generated from the corpus of the INEX collection as well as from external sources. The purpose of the experiments was to initiate tests of the principle of polyrepresentation [3] with different cognitive and functional representations of the document corpus.

The paper is structured as follows: The principle of polyrepresentation and the cognitive theory of IR interaction from which it is derived are briefly discussed as a theoretical framework for the experiments in section 2. Section 3 describes the experimental setup, and section 4 analyses the results. Section 5 gives tentative conclusions.

2 POLYREPRESENTATION

The cognitive theory of IR interaction and the principle of polyrepresentation derived from it [3] provides a theoretical background for working with different representations from several sources. In summary, it is hypothesised that overlaps between different cognitive and functional representations of both users' information needs as well as documents can be exploited for reducing the uncertainties inherent in Information Retrieval (IR), and thereby improve the performance of IR systems. Two or more different cognitive representations pointing at the same documents is regarded as multi-evidence of those documents being relevant, and suggests to apply a principle of 'intentional redundancy' [2] with the purpose of reducing the uncertainties by placing emphasis on overlaps between representations. Better results are expected when cognitively unlike representations are used, e.g., the document title (made by the author) vs. intellectually assigned descriptors from indexers.

Although cognitive theory of IR interaction and the principle of polyrepresentation is holistic in nature and amalgamates user-oriented approaches with both Boolean and best match principles it is, however, inherently *Boolean* in much of its reasoning. This is apparent in the pronounced focus on cognitive retrieval overlaps, i.e., *sets* of documents retrieved based on different cognitive representations, see, e.g., the appendix example in [3]. A little discussed, but inherent point is that the structure ensures the *quality* of the sets that are matched. But this structure does not necessarily have to be of a Boolean nature – other kinds of structure may be implemented. Such may include the probabilistic query operators in the InQuery IR system for instance as utilised by [4] to achieve various degrees of structure in queries.

Inspired by the work of Madsen and Pedersen [12] Larsen [9] proposes the idea of a polyrepresentation continuum (See Figure 1 below) as a model for discussing how structured a given implementation of polyrepresentation is.

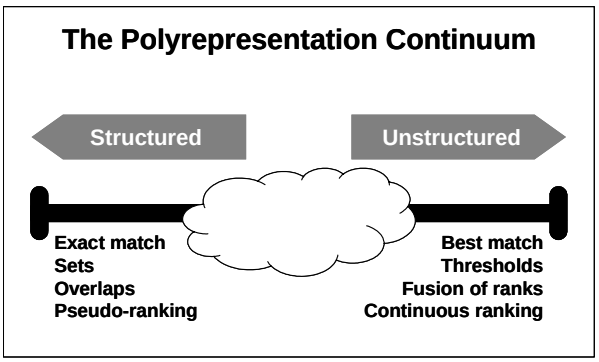

Figure 1. The polyrepresentation continuum [From 9, p. 36]

At the *structured* pole of the continuum the implementations are based on exact match principles, leading to sets of retrieved documents for each representation from which overlaps can be formed and a pseudo-ranking be constructed. At the *unstructured* pole of the continuum the implementations are based on best match principles leading to a rank of the documents that are retrieved as input for polyrepresentation. Rather than straight generation of overlaps between sets, the implementations at the unstructured pole of the polyrepresentation continuum will consist of fusing ranks to produce a final ranked output, perhaps aided by thresholds to provide the necessary quality by restricting the ranks to be fused to the top ranked documents only.

Few empirical investigations that explicitly tests the principle of polyrepresentation have been carried out so far. Larsen [8] reports a small online Boolean experiment at the structured end of the continuum. The MSc thesis of Madsen and Pedersen [12] combines a highly structured Boolean approach with probabilistic query operators in a best match system, and is as such placed closer to the middle of the continuum.

3 METHODS

The main focus of the runs submitted to INEX2003 was on obtaining functionally and cognitively different representations of the documents. Only simple fusion strategies for combining the representations were used because of lacking time to experiment with more advanced ones (See section 3.2). The runs submitted to INEX2003 were therefore close to the unstructured pole of the polyrepresentation continuum. The investigation of more advanced strategies for how to combine these in a suitable structured manner according to the principle of polyrepresentation is the subject of future work. Note that the purpose of the experiments reported in the present paper was to retrieve *whole* documents, and not document components as in most approaches in INEX.

Functionally different representations are defined as representations originating from the same cognitive agent, e.g., the article title or figure captions made by the author [3]. In relation to IR, representations are regarded as cognitively different if they originate from other cognitive agents than the author, e.g., descriptors from a thesaurus assigned intellectually to the documents, or later citations or links to the document by other authors. The corpus of the INEX test collections offers excellent opportunities for the generation of functionally different representations originating from the author because of the elaborate XML structure of the documents. In addition, a range of cognitively different representations of the documents are available because the journals in the corpus are indexed in the INSPEC database. A further opportunity offered by the INEX corpus is to exploit the references in the bibliographies to generate citation-based representations.

The InQuery IR system was used for all runs because it offers the possibility to store different representations of the documents in fields and to combine these using both Boolean and softer query operators.

3.1 Indexes and fields[1]

Two indexes were constructed, each containing three fields: one with author generated representations, one with intellectually assigned descriptors from a domain thesaurus, and one with a citation index generated from the corpus (See Figure 2 and Figure 3).

The first field consists of different types of *titles* from the documents: the article title, the section headings at all levels, and the cited titles from the bibliographies. These are either generated or selected by the author. The inclusion of section headings is inspired by the Subject Access Project (SAP) [1; 19] where section headings, figure and table captions were extracted as representations in addition to the article titles. The use of cited titles has been proposed by Kwok [6; 7], and tested by Salton and Zhang [17]. The latter experiment did not show any general gains from including cited titles. However, only those articles that were also source documents in the test collections used were included in the experiment, i.e., only a limited selection of cited titles was used in their experiments. The INEX corpus has *all* cited titles and may thus provide better results with the cited titles. The path used for extracting the cited titles was //bb/atl. This includes the titles of cited journal articles and conference papers, but not the

[1] After submission we discovered a number of errors in the indexing process. Attempts have been made to correct these, and the methods and results reported here are for the corrected runs.

titles of cited books or reports. More than 7,000 documents contained such cited titles with an average of 9.9 cited titles per document.

Titles (FLD001)
(Article title, section titles, and cited titles)
	//fm/tig/atl
	//st
	//bb/atl

Descriptors (FLD002)	Intellectually assigned descriptors
Citation index (FLD003) (Boomerang effect)	Best possible tuning with INEX2002 test collection

Figure 2. Index A (without expansion on descriptors)

The second field consists of intellectually assigned descriptors from the INSPEC thesaurus. These were available for 7,711 of the 12,107 documents in the INEX corpus. Because only relatively few descriptors are assigned to each document by the INSPEC indexers this representation contained relatively few index keys. In an effort to enlarge this representation we expanded the descriptors by adding all the synonyms (the used for (UF) relation) as well as the narrower terms (NT) from the INSPEC thesaurus. Index A contained the un-expanded descriptors (Figure 2), and Index B contained the expanded descriptors (Figure 3).

Titles (FLD001)
(Article title, section titles, and cited titles)
	/fm/tig/atl
	//st
	//bb/atl

Descriptors (FLD002) (expanded document representation)	Intellectually assigned descriptors, expanded from the INSPEC thesaurus (NT, UF)
Citation index (FLD003) (Boomerang effect)	Best possible tuning with INEX2002 test collection

Figure 3. Index B (with descriptors expanded from the thesaurus)

The third field in both indexes contained data for constructing a citation index, i.e., data to identify the references in each document. When indexed in the database documents can be retrieved that refer to (cite) a particular *seed document.* Such search strategies have shown promising results [See, e.g., 13; 14; 16], but have rarely been exploited in IR research[2]. This is probably partly due to a lack of citation data in the test collections developed in the last decade, and partly due to the lack of seed documents to represent the information need. A particular approach to identify such seeds automatically was used to construct queries for the citation index (See section 3.2). The index was constructed based on the cited titles discussed above in combination with the cited year. Because there were numerous typos etc. in the cited titles an implementation of the edit distance algorithm was used to identify variants to the same cited document[3]. 7,111 documents contained references with both cited titles and cited years. In these documents there were 70,634 unique citations after merging of variants, and these were mentioned a total of 192,881 times in the documents. The citations were represented by id-numbers to ease processing.

3.2 Queries

Only content only (CO) topics were used because only whole documents were retrieved with the tested approach.

The same queries were used for both the title field and the field containing descriptors (FLD001 and FLD002). These were constructed manually from the title elements of the CO topics translating the INEX operators into InQuery's probabilistic query operators (See Figure 4).

In order to be able to match the content of the citation index with the topics, the latter had to be translated into citations. This was done with a best match version of the so-called boomerang effect proposed in [8; 10]. In short, the boomerang effect extracted the citations from sets of documents retrieved by natural language queries from a range of functional and cognitive representations. These citations were used as seeds in a citation search that can retrieve later documents that cite the seeds. The occurrence of the citations between representations and their frequency was used to weight and select which citations to use as seeds as well as to weight the seeds in the query (See [10] for details). The boomerang effect used was the best possible tuning based on the INEX2002 test collection: citations were extracted from 8 documents resulting in 252 seed documents in average per query.

InQuery's #sum operator was used to combine the fields (See Figure 4). Only a simple strategy was used to fuse the fields because the main focus was on obtaining functionally and cognitively different representations of the documents. Therefore the runs can be characterised as being at the unstructured end of the polyrepresentation continuum. The same queries were used for index A and index B.

[2] Increasingly, web search engines exploit link data. However, there are indications that although similar in conception links and citations may be quite different in practice, see e.g., [18]. CiteSeer is an exception because it uses citations extracted from scientific papers [11].

[3] We greatly acknowledge the Department of Information Studies, University of Tampere, Finland for making the source code for this implementation available to us.

```
#sum (
#field (FLD001 #and(#1(natural language processing) (#1(human language))) #not(#1(programming language)) #not(#1(modeling language)))
#field (FLD002 #and(#1(natural language processing) (#1(human language))) #not(#1(programming language)) #not(#1(modeling language)))
#field (FLD003 #WSUM(1 3797.98 CIT_ID46361 2404.53 CIT_ID28456 1898.99 CIT_ID43757 1898.99 CIT_ID43816 1898.99 CIT_ID57141 ... )) )
```

Figure 4. Sample query (CO topic 111). Note that the citation query in FLD003 has been shortened.

3.3 Runs

The two main runs were the runs on index A and index B to study the effect of the expanded descriptors. We also did runs on the individual fields to assess their contribution to the overall result. Six runs are reported here: IndexA_run, IndexB_run, Titles_run, Descriptor_run, Descriptor_expanded_run, and Citation_index_run.

4 RESULTS

Table 1 shows the results for the strict and generalized quantification functions in inex_eval. Overall, the results display a low performance compared to the best runs in INEX2003: For instance, the highest strict AvgP value was 0.04292 for the Titles_run. The top 10 in INEX2003 was in the 0.1214-0.0664 range.

Run name	AvgP (strict)	AvgP (generalized)
IndexA_run	0.03818	0.01508
IndexB_run	0.03811	0.01510
Titles_run	0.04292	0.01550
Citation_index_run	0.03359	0.01198
Descriptor_run	0.00996	0.00724
Descriptor_expanded_run	0.00829	0.00699

Table 1. Overall results. Strict and generalized quantification functions.

Figures 5 to 7 show P-R curves for the runs. It is obvious from Figure 6 and Figure 8 as well as Table 1 that the expansion of the descriptor document representation did not improve performance; it rather decreased it slightly. The difference between the original and the expanded descriptors are not great though, and consequently the difference between the IndexA and IndexB runs are minimal (Figure 5 and Figure 7).

Figure 6 shows the performance each individual field. The un-expanded descriptors in themselves perform quite poorly (AvgP_strict = 0.00996), and the idea of expanding this representation is supported. The Titles_run have the best performance of all 6 runs (AvgP_strict = 0.04292), followed by the Citation_index_run (AvgP = 0.03359). The same patterns can be found when the results are measured with the generalized quantification function; the general level of performance is lower though.

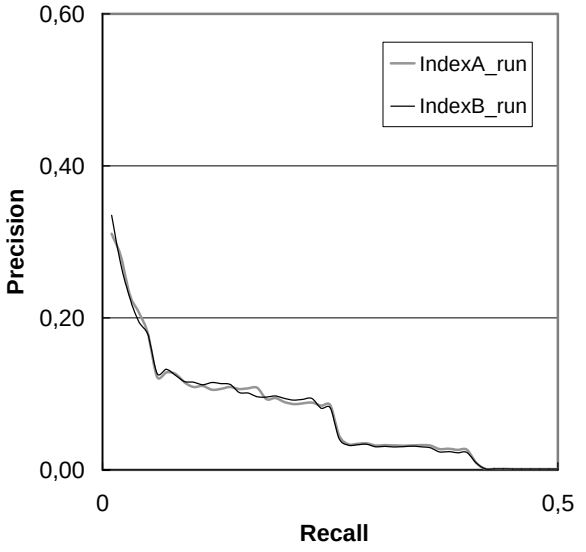

Figure 5. P-R curves for IndexA and IndexB run using the strict quantification function in inex_eval.

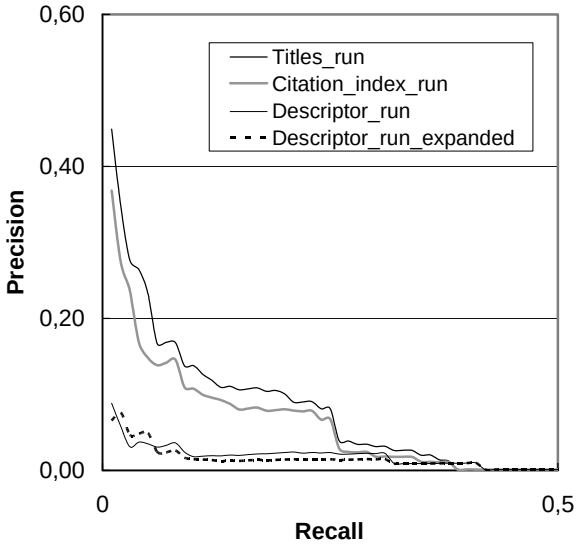

Figure 6. P-R curves for the individual fields using the strict quantification function in inex_eval.

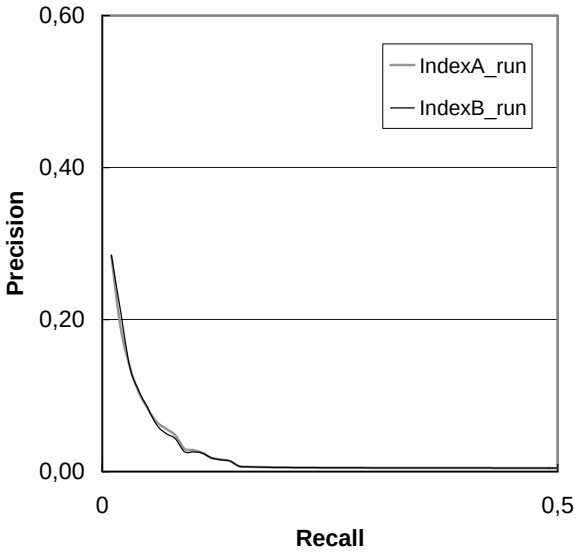

Figure 7. P-R curves for IndexA and IndexB run using the generalized quantification function in inex_eval.

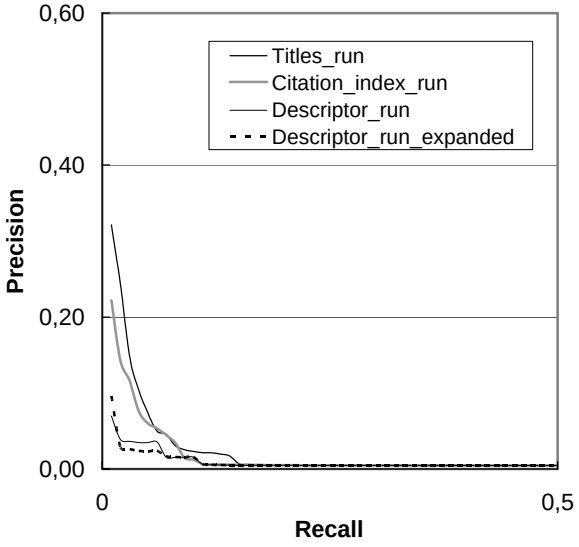

Figure 8. P-R curves for the individual fields using the generalized quantification function in inex_eval.

5 CONCLUSIONS

The overall aim of our runs submitted to INEX2003 was to work on obtaining functionally and cognitively different representations of the documents. Two of these were successful: The titles representation consisting of the article title, headings and cited titles, and the citation index, which performed fairly well. The intellectually assigned descriptors did not perform well, and it was attempted to expand these in the document representation by using the INSPEC thesaurus. This was not a success: the expansion resulted in slightly decreased performance.

Future work includes the investigation of other expansion techniques on the query side can also be implemented, e.g., similar to the ones tested in [5]. The approach tested in the runs was close to the un-structured pole of the polyrepresentation continuum. Future work also includes investigations of more advanced structured query strategies to improve the quality of the initial set used, and move the tests closer to the structured pole of the continuum.

6 ACKNOWLEDGMENTS

The InQuery IR system was provided by the Center for Intelligent Information Retrieval, University of Massachusetts, Amherst, MA, USA. The edit distance implementation used was based on the source code for the 'like' approximate string matching program [See, e.g., 15] kindly lent to us by the Department of Information Studies, University of Tampere, Finland. We wish to thank both for the use of these resources without which the research reported in this paper would not have been possible.

7 REFERENCES

1. Atherton-Cochrane, P. (1978): *Books are for use : final report of the subject access project to the Council on Library Resources*. Syracuse, N. Y.: School of Information Studies, Syracuse University. 172 p.

2. Ingwersen, P. (1994): Polyrepresentation of information needs and semantic entities : elements of a cognitive theory for information retrieval interaction. In: Croft, W. B. and van Rijsbergen, C. J. eds. *SIGIR '94 : Proceedings of the seventeenth annual international ACM-SIGIR conference on research and development in information retrieval, organised by Dublin City University, 3-6 July 1994, Dublin, Ireland*. London: Springer-Verlag, p. 101-110.

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

4. Kekäläinen, J. and Järvelin, K. (1998): The impact of query structure and query expansion on retrieval performance. In: Croft, W. B., Moffat, A., van Rijsbergen, C. J. and Zobel, J. eds. *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in*

Information Retrieval (ACM SIGIR '98), Melbourne, Australia, August 24-28, 1998. New York: ACM Press, p. 130-137.

5. Kristensen, J. (1993): Expanding end-user's query statements for free text searching with a search-aid thesaurus. *Information Processing & Management*, 29(6), 733-744.
6. Kwok, K. L. (1975): The use of titles and cited titles as document representations for automatic classification. *Information Processing & Management*, 11(8-12), 201-206.
7. Kwok, K. L. (1984): A document-document similarity measure based on cited titles and probability theory, and its application to relevance feedback retrieval. In: van Rijsbergen, C. J. ed. *Research and development in information retrieval : proceedings of the third joint BCS and ACM symposium, King's College, Cambridge, 2-6 July 1984*. Cambridge: Cambridge University Press, p. 221-231.
8. Larsen, B. (2002): Exploiting citation overlaps for information retrieval: generating a boomerang effect from the network of scientific papers. *Scientometrics*, 54(2), 155-178.
9. Larsen, B. (2004): *References and citations in automatic indexing and retrieval systems : experiments with the boomerang effect*. Copenhagen: Royal School of Library and Information Science. XIII, 297 p. ISBN: 87-7415-275-0. (PhD thesis - accepted for defence) [http://www.db.dk/blar/dissertation, visited 11-2-2004]
10. Larsen, B. and Ingwersen, P. (2002): The boomerang effect : retrieving scientific documents via the network of references and citations. In: Beaulieu, M., Baeza-Yates, R., Myaeng, S. H. and Järvelin, K. eds. *Proceedings of SIGIR 2002 : the twenty-fifth annual international ACM SIGIR conference on research and development in information retrieval, August 11-15, 2002, Tampere, Finland*. New York: ACM Press, p. 397-398. (Poster paper) [http://www.db.dk/blar (Preprint), visited 3-8-2003]
11. Lawrence, S., Giles, C. L. and Bollacker, K. D. (1999): Autonomous Citation Matching. In: Etzioni, O., Müller, J. P. and Bradshaw, J. M. eds. *AGENTS '99. Proceedings of the Third Annual Conference on Autonomous Agents, May 1-5, 1999, Seattle, WA, USA*. New York: ACM Press, p. 392-393. (Poster paper) [http://citeseer.nj.nec.com/lawrence99autonomous.html (preprint), visited 16-8-2003]
12. Madsen, M. and Pedersen, H. (2003): *Polyrepræsentation som IR metode : afprøvning af polyrepræsentationsteorien i et best match IR system [Polyrepresentation as IR method : test of the theory of polyrepresentation in a best match IR system]*. [Copenhagen]: Danmarks Biblioteksskole. 106 p.+ XLIII p. (In Danish - unpublished MLIS thesis)
13. McCain, K. W. (1989): Descriptor and citation retrieval in the medical behavioral sciences literature: Retrieval overlaps and novelty distribution. *Journal of the American Society for Information Science*, 40(2), 110-114.
14. Pao, M. L. (1993): Term and citation retrieval - a field-study. *Information Processing & Management*, 29(1), 95-112.
15. Pirkola, A., Keskustalo, H., Leppänen, E., Känsälä, A.-P. and Järvelin, K. (2002): Targeted s-gram matching : a novel n-gram matching technique for cross- and monolingual word form variants. *Information Research*, 7(2), -paper no. 126. [http://informationr.net/ir/7-2/paper126.html, visited 23-8-2003]
16. Salton, G. (1971): Automatic indexing using bibliographic citations. *Journal of Documentation*, 27(2), 98-110.
17. Salton, G. and Zhang, Y. (1986): Enhancement of text representations using related document titles. *Information Processing & Management*, 22(5), 385-394.
18. Thellwall, M. (2003): What is this link doing here : begining a fine-grained process of identifying reasons for academic hyperlink creation. *Information Research*, 8(2), paper no. 151. [http://informationr.net/ir/8-3/paper151.html, visited 8-11-2003]
19. Wormell, I. (1985): *Subject Access Project : SAP : improved subject retrieval for monographic publications*. Lund: Lund University. 141 p.