Information Processing and Management 44 (2008) 1673–1683

Contents lists available at ScienceDirect

Information Processing and Management

journal homepage: www.elsevier.com/locate/infoproman

Inter and intra-document contexts applied in polyrepresentation for best match IR

Mette Skov *, Birger Larsen, Peter Ingwersen

Information Interaction and Information Architecture, Royal School of Library and Information Science, Birketinget 6, DK-2300 Copenhagen S, Denmark

ARTICLE INFO

Article history:
Received 8 February 2008
Received in revised form 20 April 2008
Accepted 29 May 2008

Keywords:
Contextual IR
Document structure
Polyrepresentation
Cognitive overlaps

ABSTRACT

The principle of polyrepresentation offers a theoretical framework for handling multiple contexts in information retrieval (IR). This paper presents an empirical laboratory study of polyrepresentation in restricted mode of the information space with focus on inter and intra-document features. The Cystic Fibrosis test collection indexed in the best match system InQuery constitutes the experimental setting. Overlaps between five functionally and/or cognitively different document representations are identified. Supporting the principle of polyrepresentation, results show that in general overlaps generated by three or four representations of different nature have higher precision than those generated from two representations or the single fields. This result pertains to both structured and unstructured query mode in best match retrieval, however, with the latter query mode demonstrating higher performance. The retrieval overlaps containing search keys from the bibliographic references provide the best retrieval performance and minor MeSH terms the worst. 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 whilst re-ranking by citations seems less useful when integrated into polyrepresentative applications.

1. Introduction

In a cognitive approach to information interaction (Ingwersen, 1996), the different actors in the interaction processes contribute interpretations of their situations and pre-suppositions of the world as well as of the information structures and objects involved. Such manifestations of human cognition, reflection and ideas take the form of different representations, for instance, author's text, pictures, music tunes in documents, or as database designers' indexing schemes and retrieval algorithms, see information objects and information technology (IT) components of model, Fig. 1. They might also consist of a searcher's request formulation(s) and corresponding perceived work task description(s) representing his/her information requirement and problem state originating from the socio-organizational context (Ingwersen & Järvelin, 2005). Implicit or explicit patterns and evidence of seeking behaviour are similarly seen as such representations of the interaction processes (Larsen, Kekäläinen, & Ingwersen, 2006).

In this perspective the interpretations (and thus representations) found in each of the components of the model, Fig. 1, including the information interaction process itself, are seen as contextual and complementary to one another and interplay over time. This complementarity constitutes the idea behind the polyrepresentation principle.

* Corresponding author. Tel.: +45 32586066; fax: +45 32840201.
E-mail addresses: ms@db.dk (M. Skov), blar@db.dk (B. Larsen), pi@db.dk (P. Ingwersen).

0306-4573/$ - see front matter © 2008 Elsevier Ltd. All rights reserved.
doi:10.1016/j.ipm.2008.05.006

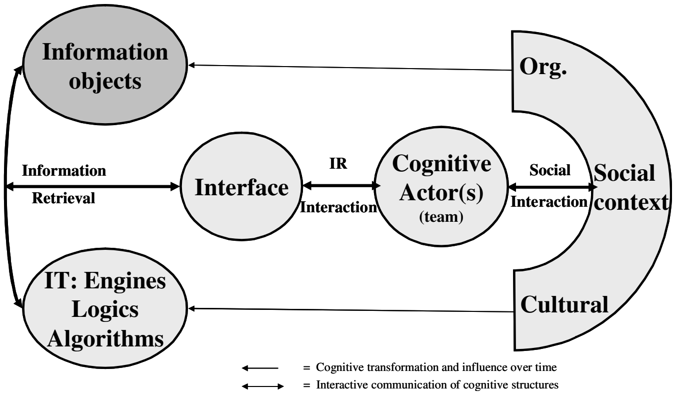
Information objects
Org.
Information
Retrieval
Interface
IR
Interaction
Cognitive Actor(s)
(team)
Social
Interaction
Social context
IT: Engines Logics Algorithms
Cultural
= Cognitive transformation and influence over time
= Interactive communication of cognitive structures

Fig. 1. Cognitive model of central components of information, IR and social interaction. The present paper focuses on different types of document representations (information objects, upper left hand). Development from Ingwersen and Järvelin (2005, p. 301).

According to Ingwersen (1992, 1996), Ingwersen and Järvelin (2005) 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 points to a set of objects in [retrieval] 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..." (Ingwersen & Järvelin, 2005, p. 208). The retrieval overlaps of sets of objects created by the divergent cognitive (and functional) representations we name 'cognitive overlaps' – Fig. 2.

Polyrepresentation distinguishes between two kinds of representations: *cognitively different* ones deriving from the interpretations by *different* actors; and *functionally different* representations that derive from the *same* actor, such as, author generated document structures, title, image features, diagram captions, and references or out-links (anchors), e.g. the information objects component (dark shading), Fig. 1, and the 'Author' part of the Venn-diagram, Fig. 2. With respect to anchor text one might argue that they inform both about the *cited* item (Brin & Page, 1998) and about the *citing* item (Schneider, 2004). Owing to their contextual properties, references are thus functionally quite different from, say, title words and other content-bearing tokens. The "selector category", Fig. 2, signifies special actors responsible for the existence and availability of the information objects, such as editors or publishers. According to the principle, good retrieval results are expected when cognitively unlike representations are used for retrieving the documents, e.g. the *same search key* found in the document title (made by the author) *and* retrieved from the intellectually assigned descriptors made by indexers, *and* found in citing documents (or in-links) made by other authors over time.

In other words, polyrepresentation is a particularly structured way of carrying out a kind of classic triangulation in the information and IT spaces and, simultaneously, allows making use of redundancy in the cognitive space of searchers. The diversification of all components in Fig. 1 including the searcher's space, information objects and IT-components and their application are founded in a theoretical IR framework and constitutes the real novelty of polyrepresentation. A more detailed

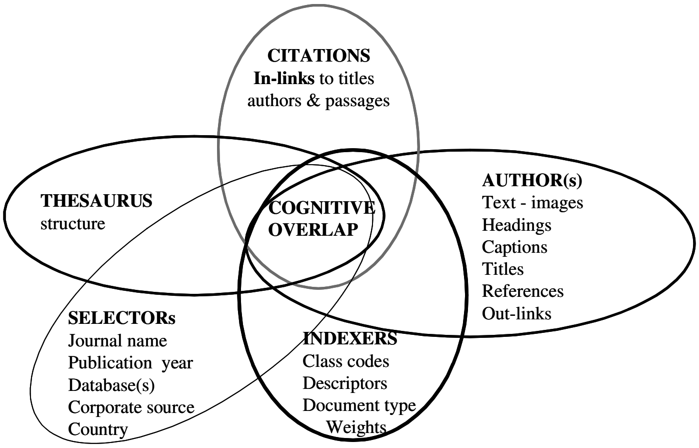

Fig. 2. The principle of polyrepresentation in academic information objects. Overlaps of information items retrieved by means of one search engine via representations of cognitively and functionally different information structures, but holding identical keys associated with one searcher statement (e.g. a work task description). Elaborated from Ingwersen (1996, p. 28), Ingwersen and Järvelin (2005, p. 207), Larsen et al. (2006, p. 149).

analysis of the scientific background underlying the principle and its applicability is provided by Ingwersen and Järvelin (2005, pp. 206–214 and pp. 342–346), Larsen et al. (2006).

Several machine learning studies have examined the use of document structure weights to learn the best combination of document structures resulting in increased retrieval performance. This work has resulted in the field-weighted BM25 ranking function by Robertson, Saragoza, and Taylor (2004). In the mentioned studies the optimal combination of document structure weights are learned by examining retrieval performance. In the present study, however, the representations are chosen and grouped based on the principle of polyrepresentation. The present study is the first of its kind to examine IR by focusing on polyrepresentation of information objects in order to explore the *best combinations* of document representations, measured according to retrieval performance. The paper investigates the intersection of search keys as found in Title/Abstract terms and phrases (made by authors), and as *identical* or similar MeSH major and minor index terms (from indexers), and as title words in documents being *cited* (as references) by the documents retrieved, as illustrated below in Fig. 3, e.g. OL 1. The references thus play the role of additional document features (context) useful to IR, as originally proposed by Garfield with respect to the creation of the citation indexes (Garfield, 1979, 1993). The present article elaborates on the results of Skov, Pedersen, Larsen, and Ingwersen (2004, 2006).

Three kinds of experiments are executed. First we test *restricted best match polyrepresentation* of the four search fields mentioned above in all their combinations. Secondly, we investigate the retrieval performance in restricted best match polyrepresentative IR of *structured queries* expanded by thesaurus structures (synonyms) according to Kekäläinen and Järvelin (1998, 2000) against unstructured (bag-of-words) queries. And thirdly, we do two re-ranking tests according to (a) obtained precision in overlaps between different representations or (b) based on citation impact. The three experiments are further explained in Section 3.1.

The paper is structured as follows: Section 2 discusses briefly previous empirical studies that are central to the understanding of polyrepresentation and associate to the present investigation. Section 3 outlines the experimental setting and places the five document representations included in the experiment in a polyrepresentation framework. Section 4 presents and discusses the results of the investigations. Section 5 concludes the article by suggesting future possible applications and research on polyrepresentation.

2. Previous empirical studies of polyrepresentative nature

2.1. Polyrepresentation of information space

The first observations of retrieval performance improvements made by polyrepresentative means were carried out by McCain (1989) and Pao (1994). However, these studies did not apply that concept. The experiments demonstrated that retrieval of documents by search keys found in titles and abstracts and by involving documents identified by a citation search strategy produces much higher odds for finding relevant documents in the constructed overlap than in each of the retrieved sets independently (Pao, 1994). The overlaps turned out to be small. In a slightly different experimental setting Christoffersen recently tried out the effectiveness of retrieval overlaps constituted by combining several databases (Christoffersen, 2004). He searched Medline, Embase and SCI in order to test the relevance proportions in any of the overlaps created online between indexers of MeSH (Medical Subject Headings in Medline), author text (Embase) and citing authors's texts (Science Citation Index). Expert relevance assessments were used. He found that "[t]he degree of overlap strongly correlates with the percentage of relevant items in a set" (p. 391). The results were statistically significant (Ingwersen & Järvelin, 2005).

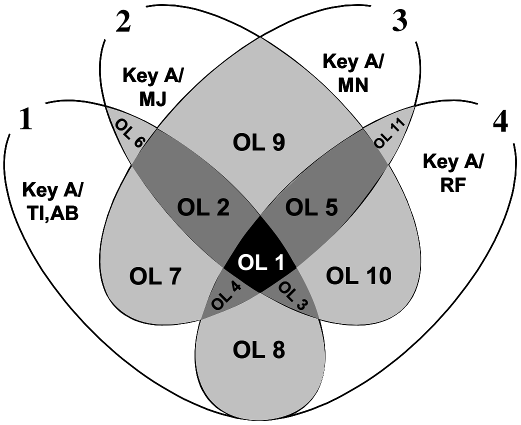

Fig. 3. Restricted polyrepresentative IR by searching key A as an intersection of all the four involved representations and all their possible combinations, isolating the overlaps. The four representations were: Major MeSH terms (MJ), References (RF), Title/Abstract (TI/AB) and Minor MeSH terms (MN). Elaboration from Ingwersen and Järvelin (2005, p. 208).

Also in recent investigations Larsen (2004) tested the retrieval performances of a variety of inter and intra-document features in the information space, the focus component, Fig. 1. He used different document representations from the INEX test collection[1], as well as added thesaurus and uncontrolled terms assigned by INSPEC indexers. Finally, *citation cycling* strategies, i.e. backward chaining followed by forward citation chaining were used, named the *Boomerang effect* (Larsen et al., 2006). The purpose of the Boomerang effect was to exploit the potentially high performance offered by incorporating link or citation information, and at the same time allow for queries formulated in natural language.

The best precision result was achieved by functionally different document representations, such as article titles, section headings and the *cited titles* in the references. Reasonable effectiveness was obtained by combining those representations and intersecting them with indexer descriptors and the citation cycling based Boomerang effect (Larsen, 2004). Unstructured queries were used, and the frequency of documents cited was part of the weighting in the citation search. The Boomerang effect was compared (1) to best match bag-of-words used for each representation separately and fused in a relaxed polyrepresentative modus and (2) to a common bag-of-words based baseline. The results showed that the Boomerang effect on average did not decrease performance, but relaxed polyrepresentation was slightly better. However, the common bag-of-words baseline obtained the best overall performance (Larsen et al., 2006, p. 155).

2.2. *Polyrepresentation of users' cognitive space*

As noted above the entire principle of polyrepresentation does not rely on *one* request formulation from a searcher. It assumes that functionally different representations of the searcher's cognitive space come into play too, such as current knowledge state of domain, request formulation or work task perception (see Ingwersen, 1996). Again, one may regard such polyrepresentations as restricted when using each representation separately to query the retrieval engine and isolating the output to be measured for performance to the inner overlap by intersection; or as more relaxed if loosely combined by a union of the representations with term weights, as in recent investigations by Kelly, Dollu, and Xin (2005), relating back to Belkin et al.'s multi-query version experiments (Belkin, Kantor, Fox, & Shaw, 1995).

Kelly et al. (2005) combined different searcher statements of a single information need with statements captured from each user via a structured interface: current knowledge state of the request; reason for the request; and additional known concepts. The experiments were carried out within the TREC HARD track using the Lemur IR Toolkit. A baseline was constructed from the TREC topic title and description elements combined into a single query. The investigation's experimental (single) runs consisted of the baseline query terms + different combinations of relevance feedback modes + the polyrepresentative combination from the searcher's space. Whereas the baseline alone yielded on average 9.3 words, the polyrepresentation captured from the searchers yielded more than 25 different words on average. This result stresses the usefulness of attempting the extractions of searcher statements, in particular if initial requests (the TREC topic title) are very short statements. When combining the polyrepresentative searcher statements, Kelly and colleagues did not retrieve documents by each statement separately later to be intersected in order to define overlaps of documents. Instead, they applied the statements in union with weights for term repetition (term overlap). This is thus a relaxed approach to polyrepresentation. Their retrieval performance results are highly promising from a polyrepresentation point of view, since all polyrepresentation combinations yielded statistical significant performance improvements over the baseline (Kelly et al., 2005). Also, they found a significant correlation between query length and performance.

2.3. *Polyrepresentation of retrieval engines and the interaction process*

Lund, Schneider, and Ingwersen (2006) carried out polyrepresentation studies of the IT system component (lower left model component, Fig. 1). They examined the retrieval results from the 12 most effective TREC 5 search engines. In Lund et al.'s study the involved search engines illustrate *cognitively* different representations of IR system settings when they follow different retrieval principles, and *functional* difference when they are versions of the same principle. Their initial results demonstrate that when combining the search engines according to restricted polyrepresentation principles, the performance in terms of recall and precision depends on how many relevant documents potentially exist in the search task. The more relevant documents in the topics over the engines in all combinations the higher the precision (Lund et al., 2006). This result implies that in data fusion based on polyrepresentation principles search topics should contain 'substantial numbers' of relevant documents in order to function properly. According to Larsen et al. (2006, p. 154), Lund et al.'s results indicate that a combination of 3–5 systems may perform better than combinations of, for instance, 12 different IR engines, owing to the involvement of too many functionally alike and/or badly performing IR systems in the fusion.

Secondly, for the restricted combinations of the four best performing but logically different engines, for all performance indicators, the combined fusion outperformed the single systems. In support of the polyrepresentation principle Lund et al.'s results demonstrate with statistical significance that when *cognitively different systems* are combined the precision is higher than when only functionally different systems are fused – provided that they are of equal performance. These results coincide with automatic IR data fusion experimental findings that smaller retrieval overlaps commonly imply better

[1] The initiative for the evaluation of XML retrieval (INEX) test collections from 2002 to 2005 are based full text academic articles from the journals of the IEEE Computer Society (see Lalmas & Tombros, 2007).

Table 1
Summary figures for the 29 topics from the Cystic Fibrosis test collection used in this study

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
Field	Number of records	Number of words
Title (TI)	1239	12,811
Abstract (AB)	1239	508,439
Major MeSH terms (MJ)	1236	3469
Minor MeSH terms (MN)	1239	13,557
Author (AU)	1209	3373
References (RF)	1195	27,043

performance of the fusion over the single systems involved (Kwong & Kantor, 2000; Wu & McClean, 2006). Further experimental results on data fusion following 'relaxed' polyrepresentative principles are underway at present.

Looking into the interaction process itself (the horizontal line, Fig. 1) White has recently investigated empirically the principle of polyrepresentation applied to *interface functionality* (2006) and *implicit relevance feedback* algorithms for interactive IR (White, Ruthven, Jose, & van Rijsbergen, 2005). White proposes to apply "content-rich search interfaces that implement an aspect of polyrepresentation theory, and are capable of displaying multiple representations of the retrieved documents simultaneously in the results interface" (White, 2006, p. 1). Three 'simulated search' scenarios (Borlund, 2003) were tested for retrieval performance by means of searcher simulations of all possible combinations of representations and paths available. The best performing combination of representations consisted of document title, its query-biased summary and summary sentence in context.

3. Testing inter- and intra-document features in polyrepresentation – the experimental setting

The present project and experimental setting was inspired by and elaborated on the Larsen observations (2002, 2004), and the Pao experiments (1994) discussed above. The setting included selected elements of polyrepresentation of the information space in a best match setting. The Cystic Fibrosis test collection (Shaw, Wood, Wood, & Tibbo, 1991) and the InQuery[2] retrieval system (version 3.1) were used in the experiment. The Cystic Fibrosis test collection contains 1239 document surrogates of documents published in 1974–1979. The surrogates consist of title, abstract, major and minor Medical Subject Headings (MeSH) from Medline,[3] as well the reference list and subsequently citing documents in condensed form. The Cystic Fibrosis collection originally contains 100 topics from which we randomly selected 29 for our test purpose. We chose to reduce the number of topics because the processing of each topic in the experiment was quite demanding. The topic processing was demanding primarily due to two manual elements of the process (the query expansion of search keys in MeSH thesaurus and the chosen approach to include bibliographic references as a representation) further described in Section 3.1. The relevance assessments in the test collection are given on a graded scale (retrieved documents were judged highly relevant, marginally relevant or not relevant) by four assessors. The four assessors include three medical subject experts and one medical bibliographer. Assessments provided by the three medical subject experts were used as a conglomeration. In the experiments a document is graded highly relevant if just one of three medical expert assessors have judged the document highly relevant. Likewise with documents assessed marginally relevant. Assessments by the medical bibliographer were not included since less agreement was found between relevance evaluations made by the medical bibliographer and the subject experts than between the three subject experts (Shaw et al., 1991, p. 354). Table 1 shows summary figures for the 29 topics from the Cystic Fibrosis test collection we used in the present study.

The Cystic Fibrosis test collection is small compared to common TREC news test collections or even the INEX collection of IEEE CS journal papers. However, it is ideally suited for testing inter and intra-document features in polyrepresentation as it contains both cognitively and functionally different features. *Functionally* different features were applied in the form of titles (TI) combined with abstracts (AB) and references (RF), all representing the author, combined with Medical Subject Headings (MeSH) representing the indexer as a cognitively different actor. MeSH is a controlled vocabulary used for indexing, and both major (MJ) and minor (MN) subjects of the document are represented. The TREC and INEX test collections do not contain descriptors or citations to the indexed documents and are hence less suitable for polyrepresentation experiments. Although

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

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

Table 2
Recall and precision for the 15 overlaps of restricted polyrepresentation

Overlap	Natural language queries			Highly structured queries				
	# doc	P all rel (%)	R all rel (%)	# doc	P all rel (%)	P highly rel (%)	R all rel (%)	R highly rel (%)
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)	16,201	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

Based on Skov et al. (2006). OL = overlap, P = precision, R = recall, # doc = no. of retrieved documents for all 29 test requests. Recall = 0 signifies values below 0.005%.

the Cystic Fibrosis collection contains all the bibliographic references given to previous work, they are unfortunately presented in a very short form that makes searching their titles impossible, e.g. as “Irvine WJ Lancet 2 163 970”. This form also pertains to the *citations* given to the articles from their publication year (1974–1979) until 1987. There exists a definitive need for a large-scale journal article test collection with full text, human indexing and references as well as citations for an extensive time period. Elsevier’s SCOPUS[4] might be used as a basis for creating such a collection.

3.1. The experimental setting and operations

Three kinds of experiments are executed based on the principle of polyrepresentation:

(1) In the first experiment, we test *restricted best match polyrepresentation* of the four search fields mentioned above in all their combinations. ‘Restricted’ polyrepresentation implies that a document is isolated to only one overlap, e.g. OL 1, and thus cannot be found in other overlaps, Fig. 3. This is achieved by means of the quorum searching technique (Cleverdon, 1984). Within an overlap documents are ranked using a standard retrieval engine. That is, in the combination of field 2 (RF) and field 3 (TI/AB) only the retrieval result for documents located and ranked in OL 8 is measured.

(2) Secondly, the study investigates the retrieval performance in restricted best match polyrepresentative IR of *structured queries* expanded by thesaurus structures (synonyms) according to Kekäläinen and Järvelin (1998, 2000) against unstructured (bag-of-words) queries. The study thus focuses on the Author, Indexer and Thesaurus spaces of the diagram Fig. 2. The motivation behind the two first experiments was to examine whether the overall principle of polyrepresentation was supported or not. That is, that a high number of different representations pointing towards a document is likely to be an indicator of it being relevant (Ingwersen, 1996). Retrieval performance of both structured and unstructured queries was compared to evaluate the impact of query structure. The results of the first two experiments were used as input in re-ranking tests described next.

(3) Third, 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 produced by other authors citing the documents over time can improve performance. Thus, the first re-ranking test is based on the findings of the two first experiments. Whereas in the second re-ranking test a fifth document representation, citations, is included. Citations are cognitively different from the four other representations and therefore interesting to include in polyrepresentative study. In the re-ranking tests the polyrepresentation modus is less restricted as the documents in all overlaps are fused into a single, continuous ranking.[5] This is achieved by combining, in a single query, the isolation of all overlaps and at the same time giving individual weights to each overlap. Here, higher weights from six different weighting schemes were given to documents in overlaps that as a whole showed high precision in the first and second experiment. In the aforementioned combination between the RF and TI/AB fields all the overlaps are included, except OL 9 and the single remaining fields MJ and MN.

[4] <http://www.scopus.com>.

[5] ‘Relaxed’ polyrepresentation implies that the same document can be found in *several* overlaps, depending on the combination of searchable fields. The sum of a document’s retrieval status value (RSV) will then determine its ranking, as in CombSUM in data fusion (Fox & Shaw, 1994).

Table 3
Cumulative gain: runs 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

DCV = 30. Twenty nine requests, moderately restricted polyrepresentation principle.
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.

Before we conducted the experiments we merged two representations and decided how to include the reference representation. First, the representations TI and AB were merged and searched as one representation (TI/AB), owing to the very large overlap (80%) the two fields in between. Combining the four representations (TI/AB, MJ, MN and RF) resulted in 15 overlaps[6] (Table 2). The restricted polyrepresentation search principle used in the project with respect to overlaps is shown in Fig. 3 and demonstrated below in terms of a search string from InQuery applying the quorum principles.

Secondly, a straightforward way to include references (RF) as a document representation in a polyrepresentation search is to perform a topic search in the natural language title included in the bibliographic reference in the reference list of the documents. As stated above, this option was not possible in the test collection. Therefore, an alternative contextual approach was used in order to include references as a document representation despite the missing title. Inspired by an earlier study of polyrepresentation (Larsen, 2002) references were included by means of a subject search executed for each request in science citation index (SCI). The aim of this subject search was to locate seed documents in the citation database.[7] The *references* in the retrieved documents were ranked using the Dialog RANK command on the cited reference field. This command ranks the cited references in the retrieved documents by frequency. For each request the top-three cited references were selected as seed documents.[8] The three seed documents were then used as input in a (RF) search in the test collection and provided the substitute retrieval of the title words of the references. This way of deriving at polyrepresentation overlaps by keys found in RF from context and the three other fields signifies a conservative means of polyrepresentation.

Twenty nine topics and two types of queries were tested in InQuery: 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 and Järvelin's approach to query structuring (Kekäläinen & Järvelin, 1998, 2000): first, queries were parsed for noun-phrases, shown in the example below in double quotes, e.g. "hepatic complications". Secondly, stop words were removed, and finally the remaining search keys were expanded manually using the MeSH thesaurus to increase recall. In the manual query expansion narrower terms and synonym terms were added from the MeSH thesaurus. The process of query expansion could be automated in future larger scale experiments. In accordance with Kekäläinen and Järvelin (1998) InQuery's Boolean operators were used to express relations between search terms in the highly structured queries. Using highly structured queries tend to ensure that documents identified in an overlap have identical or synonym search terms present from all the representation searched agreeing with the principle of polyrepresention (see also Section 4.2). The following example of a highly structured query is based on the test request: *What are the hepatic complications of cystic fibrosis?* In the query example below the query terms in italics are expanded terms from MeSH. It illustrates a quorum OL 2 search where the unique overlap between TI/AB, MJ, MN is identified (Fig. 3 and Table 2) by isolating it from the documents that also contain the search keys in their RF (by quorum NOT-logic excluding the documents already retrieved in OL 1):

[6] We use the term 'overlap' even though the documents in overlaps 12–15 were retrieved from one representation only, with all already retrieved documents extracted from their results.

[7] The subject search was limited to publication years 1974–1984. This is an extension compared to the publication year in the test collection. The extension resulted in a larger search result and hence a more robust basis for identifying seed documents.

[8] A seed document should have received at least three citations in SCI. In addition, a seed document must be published in 1976 or earlier in order to have received citations, as the publication window in the collection is 1974–1979. "The criteria, that a seed document must be published in 1976 or earlier, is set to ensure that as many documents in the test collection as possible should have had the chance to cite the seed documents. Seed document published later than 1976 could not have been cited by more than half of the documents in the test collection (documents published 1974–1976)".

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 sample expressed in the query language of InQuery:[9]

#q = #bandnot (#band (#field (TI/AB #syn (hepatic *liver hepatectomy* #1 ("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 an independent test variable provides information on the impact of query structure when applying polyrepresentation in a best match setting.

In the last two re-ranking experiments the following approach was used. For both experiments, the search results from the 15 overlaps (all the retrieved documents) were merged into *one* ranked list of documents (run 1, Table 3). Since each document only appears once in an originally retrieved overlap (because of the restricted polyrepresentation), the original InQuery RSV decided this run's final ranking, including documents from less well performing overlaps (like OL 2 and OL 6–9). This final ranking thus signifies a moderately restricted kind of polyrepresentation.

The first and primary re-ranking test (runs 2–4) was based on the precision obtained in the structured queries in the 15 overlaps (the right-hand side, Table 2) by assigning weights to the documents according to (a) precision obtained in those queries and (b) precision obtained retrieving highly relevant documents (column 7, Table 2). The following weights were applied (in run 1 no weights were applied). In run 2 weight 100 was assigned the documents in the four overlaps with the highest precision in structured retrieval (OL 1, 3, 4, and 5). In run 2 no weights were assigned to documents in OL 2, as precision obtained in this overlap is considerably lower than in the before mentioned overlaps with three and four representations (column 7, Table 2). This weighting is intended to reduce the impact of the documents placed in the less well performing overlaps when all the overlaps are merged into one ranked list of documents. It boosts the best polyrepresentative combinations as if the selected overlap documents are repeatedly retrieved 100 times. In addition, run 2 illustrates weights assigned to cognitively and functionally different fields. In run 3 weight 100 was applied as in run 2 plus that weight 50 was applied to overlaps with medium precision (OL 2, 6, 8, and 10, Table 2). This run also introduces documents from the inner overlaps retrieved simultaneously by structured queries from only two fields, regardless types of overlaps. In run 4 weight 100 was applied to overlap 1 (consisting of four different representations), weight 66 was applied to OL 2–5 (consisting of three different representations), and weight 33 was applied to OL 6–11 (consisting of two different representations). This modus of weighting illustrates an automatic assignment of weights according to number of overlapping fields. The scaling of the weights is not learned from the experiments. Instead, the weight 100 is an absolute weight and the other weights are decided proportionally. However, more tests could reveal a more optimal assignment of weights.

The second re-ranking test (runs 5 and 6) was based on *citation impact*, i.e. applying the number of received citations as an indication of quality. Citations (in-links) represent the citing author's cognitive structures and their interpretations of the work cited. In runs 5 and 6 all documents included in the search result had received at least one or three citations, respectively.

We measured the performance of individual overlaps using straight recall and precision. The performance of the re-ranking runs were evaluated by cumulative gain (CG) (Järvelin & Kekäläinen, 2002).

4. Results and discussion

The tripartite relevance assessments provided in the test collection made it possible to investigate the retrieval performance for (a) all relevant documents (marginally relevant and highly relevant documents) and (b) highly relevant documents across unstructured and structured queries, respectively.

4.1. Retrieval performance of number and types of representations

The results show that, with a few exceptions, overlaps generated by three or four representations have higher precision than those generated from two overlaps, or from the lists of unique remaining documents retrieved by the individual search fields, as a result of the quorum searching. This observation concerns both the structured and the unstructured natural language search mode (Table 2). These findings support the overall principle of polyrepresentation suggesting that a high number of different representations pointing towards a document are likely to be an indicator it being highly relevant (Ingwersen, 1996).

However, the results also clearly demonstrate that the *types of representations* producing the best performing overlaps are central in polyrepresentation of information space. Observe, for instance, OL 3 in natural language querying ($P = 48\%$) and in

[9] The following InQuery operators were used:
- *Bandnot* and *band* in the InQuery query language are equivalent to Boolean NOT and AND, respectively.
- The keys within a *#syn operator* are treated as instances of the same term.
- The keys within a *#1 operator* must be found within one word of each other.

structured querying for all relevant ($P = 79\%$) – the highest performance achieved. These results are due to the *cognitive difference* between fields. The combinations of TI/AB, MJ and RF seem the most powerful ones, in particular those including RF (OL 1; OL 3–5; OL 10). On the other hand, precision drops when MN terms contribute to an overlap (OL 7, 11, and 14). The latter observation probably derives from the higher specificity of the MN terms, compared to the small size of the test collection, and hence its low performance. The same phenomenon of precision drop can be observed with respect to the combinations of TI/AB terms. For both MN and TI/AB their individual retrieval results (and their combination; OL 7) are very poor, Table 2. These findings stress the importance of using representations that are both cognitively dissimilar (e.g. RF vs. MJ MeSH-headings: OL 1, 3, 10) and (strongly) functionally different (e.g. TI/AB vs. RF). Exactly because of their different functionality, owing to their temporal nature and serving as additional contextual "descriptors" of the article contents, the references become central to improving IR performance, as originally intended by Garfield (1979) when devising the citation indexes.

One should also note that MJ combined with TI/AB provide better precision results than each field separately (OL 1–3 and OL 6). Table 2 further shows the impossibility of applying mean average precision (MAP) performance measures owing to the diversity of number of retrieved documents.

These observations are valid across precision for 'all relevant' as well as 'highly relevant' *P*-measures.

4.2. Impact of query structure

For all 15 overlaps highly structured queries result in higher precision for 'all relevant' documents when comparing to queries in natural language (Table 2, column 3 and 4). This supports the findings of Kekäläinen & Järvelin (1998). 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 as such to be present. This is because the NL queries are treated as best match queries by InQuery. As a consequence the retrieved sets and overlaps between them include document representations with no or little relation to the information need. Therefore, it seems that polyrepresentation in the true sense of the concept is less likely to be achieved with weakly structured queries in natural language.

4.3. The re-ranking experiments in restricted polyrepresentation

In large scale databases ranking of highly relevant documents at the top of the search result is important to end users. For this reason we measure performance using cumulative gain (CG) rather than mean average precision as more emphasis is placed on the top-ranks in CG (Järvelin & Kekäläinen, 2002). Table 3 demonstrates CG performance with document cut-off value (DCV) = 30.

To decide whether differences in the CG figures are statistically significant we ran the non-parametric Friedman test, which is based on ranks. The test was based on normalised average CG vectors as suggested by Järvelin & Kekäläinen (2002). 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, $p < 0.05$). In run 1 only the RSV given each document by InQuery determined the ranking. Documents from overlaps other than the inner one might thus be ranked high on the final output list. Results in Table 3 also indicate that *run 2*, boosting the high-precision inner overlap documents retrieved by 3–4 representation fields, constantly performs better than runs 3–6 over all DCV points. However, the differences are not statistical significant (probably partly caused by the small test collection). Comparing weighting according to polyrepresentation (runs 2–4) and InQuery's own weighting of natural language queries (bag-of-words) shows no statistically significant difference. Again run 2, however, performs better than the InQuery baseline from DCV = 5 documents over all 29 requests. This suggests that weighting according to the principle of polyrepresentation performs *at least as well* as InQuery's standard bag-of-words weighting with respect to pushing relevant documents up the ranking list.

4.3.1. Ranking by citations

Results in Table 3 indicate a slight decrease in retrieval performance when applying citations (runs 5–6) as an indication of quality compared to the best run of weighted overlaps (run 2). However, the Friedman test showed no statistical significance between the bag-of-words baseline and the two runs including citations. Table 3 further demonstrates that InQuery's baseline results in top 15 outperform runs 1 and 3–6. The Cystic Fibrosis test collection does not include citation titles, but like references only a very short form is presented, and therefore citations are not used according to the principle of polyrepresentation (Ingwersen, 1996) in the present study. In addition, because the volume of citations is small, re-ranking based on these is problematic and should be tested with much larger test collections in citation rich domains.

5. Discussion of future work and conclusions

With very few exceptions the presented experiments support the principle of polyrepresentation suggesting that a substantial number of cognitively and functionally different document representations that simultaneously point towards a document is likely to be an indicator of it being (highly) relevant, thus providing improved performance in terms of precision. The main finding suggests that successful polyrepresentative IR by combining document search fields seems to depend on (a) applying document representations as cognitively different as possible and (b) excluding poorly performing fields in the combinations, as observed in data fusion of retrieval engines (Kwong & Kantor, 2000; Wu & McClean, 2006). Such a field, as the MN MeSH field in the present tests, decreases the combined performance in natural language as well as highly structured query mode. It also seems to degrade the ability of the fusion to rank relevant documents high on the output lists.

Because of the Boolean nature of polyrepresentation, a strongly structured query language appears to be 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 overlapping fields with respect to alternative re-ranking weighting methods. Polyrepresentation equals ranking performance confronting InQuery's bag-of-words weighting scores. Results from re-ranking test based on citation impact indicate a slight decrease in performance. But the citation volume applied to the test bed was quite small.

The present experiments test the polyrepresentation principle in a conservative manner, i.e. in its restricted form. 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. An alternative polyrepresentation approach is following a more relaxed mode. This implies to allow documents to appear (be repeated) in the overlaps and to include the sub-overlaps in the final rank list, as done in common data fusion of retrieval engines. The advantage is that the repetition allows for aggregating their RSV scores in a variety of ways that can be tested. One kind of aggregation corresponds to the CombSUM weighting scheme in data fusion (Fox & Shaw, 1994) and can be illustrated by Fig. 3. The inner cognitive retrieval overlap of the four fields (OL 1) contains (few) documents for which their four assigned RSV will be aggregated. The documents found in the next level of overlaps (OL 2–5) then become mixed with the OL 1 list of documents according to their CombSUM scores, and so on until a DCV has been reached.

The Cystic Fibrosis collection mirrors scientific open access repositories that include 'non-author' derived representations, such as index keywords, citation networks, the possibility of ontology and other added metadata. In more general and often heterogeneous collections, such as Web 2.0, enriched OPACs (Library 2.0), the semantic Web, several alternative sources of "non-author"-based representations can be acquired and used. Among such are anchor texts or the entire text and other symbols or signs in Web pages linking to a specific web page, various forms of added *context* like peer-to-peer structures, social tags, reviews and recommendations of a web page or site. With an increased inter-(con)textuality of produced documents the amount and weight of non-author representations will increase and become feasible from a polyrepresentative perspective. An important issue is that the principle of polyrepresentation should be tested on larger data sets. One possibility is the large collection of Medline records used by the TREC genomics track or the IEEE INEX collection. Also 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. In Google Scholar tests of the citation-based ranking might likewise yield knowledge of how to explore polyrepresentation principles in IR.

References

Belkin, N. J., Kantor, P., Fox, E., & Shaw, J. A. (1995). Combining the evidence of multiple query representations for information retrieval. *Information Processing & Management, 31*, 431–448.

Borlund, P. (2003). The IIR evaluation model: A framework for evaluation of interactive information retrieval systems. *Information Research* 8(3), Paper no. 152. <http://www.informationr.net/ir/8-3/paper152.html> Cited 06.02.08.

Brin, S., & Page, L. (1998). The anatomy of a large-scale hypertextual web search engine. *Computer Networks and ISDN Systems, 30*(1–7), 107–117.

Christoffersen, M. (2004). Identifying core documents with a multiple evidence relevance filter. *Scientometrics, 61*(3), 385–394.

Cleverdon, C. W. (1984). Optimizing convenient online access to bibliographic databases. *Information Services and Use, 4*(1-2), 37–47.

Fox, E. A. & Shaw, J. A. (1994). Combination of multiple searches. In *Proceedings of the TREC 3* (pp. 500–215). NIST Special Publications.

Garfield, E. (1979). *Citation indexing: Its theory and application in science, technology and humanities*. New York, NY: Wiley [Reprinted by ISI Press, 1983].

Garfield, E. (1993). KeyWord Okys (TM): Algorithmic derivative indexing. *Journal of the American Society for Information Science, 44*(5), 298–299.

Ingwersen, P. (1992). *Information retrieval interaction*. London: Taylor Graham.

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

Ingwersen, P., & Järvelin, K. (2005). *The turn: Integration of information seeking and retrieval in context*. Dordrecht: Springer.

Järvelin, K., & Kekäläinen, J. (2002). Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems, 20*(4), 422–446.

Kekäläinen, J., & Järvelin, K. (1998). The impact of query structure and query expansion on retrieval performance. In W. B. Croft, A. Moffat, C. J. van Rijsbergen, R. Wilkinson, & J. Zobel (Eds.), *Proceedings of the 21st annual international ACM SIGIR conference on research and development in information retrieval* (pp. 130–137). Melbourne, Australia, New York, NY: ACM Press.

Kekäläinen, J., & Järvelin, K. (2000). The co-effects of query structure and expansion on retrieval performance in probabilistic text retrieval. *Information Retrieval, 1*(4), 329–344.

Kelly, D., Dollu, V.D., & Xin Fu. (2005). 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* (pp. 457–464). New York NY: ACM Press.

Kwong, B. Ng., & Kantor, P. (2000). Predicting the effectiveness of naïve data fusion on the basis of systems characteristics. *Journal of American Society for Information Science, 51*(13), 1177–1189.

Lalmas, M., & Tombros, A. (2007). Evaluating XML retrieval effectiveness at INEX. *SIGIR Forum, 41*(1), 40–57.

Larsen, B. (2002). Exploiting citation overlaps for information retrieval: Generating a boomerang effect from the network of scientific papers. *Scientometrics, 54*(2), 155–178.

Larsen, B. (2004). *References and citations in automatic indexing and retrieval systems: Experiments with the Boomerang effect*. Ph.D. Thesis. Copenhagen, DK: The Royal School of LIS. <http://www.db.dk/blar/dissertation> Cited 06.02.08.

Larsen, B., Kekäläinen, J., & Ingwersen, P. (2006). The polyrepresentation continuum in IR. In I. Ruthven et al. (Eds.), *Information interaction in context: International symposium on information interaction in context: IIIiX 2006: Copenhagen, Denmark, 18–20 October, 2006: Proceedings* (pp. 148–162). Copenhagen: Department of Information Studies, Royal School of Library and Information Science.

Lund, B. R., Schneider, J. W., & Ingwersen, P. (2006). Impact of relevance intensity in test topics on IR performance in polyrepresentative exploratory search systems. In Ryen White, G. Muresan, & G. Marchionini (Eds.), *Evaluating exploratory search systems, proceedings of the SIGIR 2006 EESS workshop* (pp. 42–46). <http://www.research.microsoft.com/%7Eryenw/eess>.

McCain, K. W. (1989). Descriptor and citation retrieval in the medicine behavioral sciences literature: Retrieval overlaps and novelty distribution. *Journal of the American Society for Information Science, 40*, 110–114.

Pao, M. (1994). Relevance odds of retrieval overlaps from seven search fields. *Information Processing & Management, 30*(3), 305–314.

Robertson, S., Saragoza, H., & Taylor, M. (2004). Simple BM25 extension to multiple weighted fields. In *Proceedings of the 13th ACM international conference on information and knowledge management, November 8–13, 2004, Washington, DC, USA*.

Schneider, J. W. (2004). *Verification of bibliometric methods' applicability for Thesaurus construction*. Ph.D. Thesis. Copenhagen, DK: Royal School of Library and Information Science. <http://www.biblis.db.dk/uhtbin/hyperion.exe/db.jessch04> Cited 06.02.08.

Shaw, W. M., Wood, J. B., Wood, R. E., & Tibbo, T. R. (1991). The cystic fibrosis database: Content and research opportunities. *Library and Information Science Research, 13*, 347–366.

Skov, M., Pedersen, H., Larsen, B., & Ingwersen, P. (2004). Testing the principle of polyrepresentation. In B. Larsen (Ed.). *Information retrieval in context, proceedings of the SIGIR 2004 IRiX workshop*. Sheffield, UK, July 2004 (pp. 47–49). <http://www.ir.dcs.gla.ac.uk/context/IRinContext_WorkshopNotes_SIGIR2004.pdf> Cited 06.02.08.

Skov, M., Larsen, B., & Ingwersen, P. (2006). Inter and intra-document contexts applied in polyrepresentation. In I. Ruthven et al. (Eds.), *Information interaction in context: International symposium on information interaction in context: IIIiX 2006: Copenhagen, Denmark, 18–20 October, 2006: Proceedings* (pp. 163–170). Copenhagen: Department of Information Studies, Royal School of Library and Information Science.

White, R. W. (2006). Using searcher simulations to redesign a polyrepresentative implicit feedback interface. *Information Processing & Management, 42*(5), 1185–1202.

White, R. W., Ruthven, I., Jose, J. M., & van Rijsbergen, C. J. (2005). Evaluating implicit feedback models using searcher simulations. *ACM Transactions on Information Systems, 23*(3), 325–361.

Wu, S., & McClean, S. (2006). Improving high accuracy retrieval by eliminating the uneven correlation effect in data fusion. *Journal of the American Society for Information Science and Technology, 57*(14), 1962–1973.

Mette Skov is Ph.D. student at the Department of Information Interaction and Information Architecture, Royal School of Library and Information Science, Denmark. The topic of her Ph.D. work is access to cultural heritage collections and information seeking behaviour of virtual museum visitors.

Birger Larsen is Associate Professor at the Department of Information Interaction and Information Architecture, Royal School of Library and Information Science in Copenhagen, Denmark. His research interests include information retrieval (IR), exploiting document structure and context in IR, XML IR and user interaction, Informetrics/Bibliometrics, citation analysis and research evaluation.

Peter Ingwersen is Full Professor at the Royal School of Library and Information Science, Denmark, Ph.D. in 1991. Research areas: interactive IR; evaluation methods for work task-based IR; informetrics-scientometrics & webometrics. He has published several books, and more than 70 journal articles and conference papers, in addition to editing work.