In: Spink, A. & Cole, C., (eds.), *New Directions in Cognitive IR Theory.* Springer, 2005: 43-60.

Chapter 3. Cognitive Overlaps along the Polyrepresentation Continuum

Birger Larsen and Peter Ingwersen
Royal School of Library and Information Science

1. Introduction

The principle of polyrepresentation, proposed more than 10 years ago, provides a holistic and explicitly cognitive framework for understanding the processes involved in Information Retrieval (IR) (Ingwersen, 1994, 1996, 2001). While readily applicable to the phenomena encountered in mainstream algorithmic IR research, the main strength of the principle is that it can also be applied simultaneously to the cognitive space of the user – thus integrating the two perspectives into one coherent cognitive framework. The main idea in the principle is that document overlaps generated from representations of different cognitive and functional origins can improve performance in IR systems. This kind of overlaps we entitle “cognitive overlaps”.

This chapter outlines the principle of polyrepresentation with a focus on the representations involved. The potentials and problems of the principle are discussed in the light of recent empirical studies, and challenges and opportunities for future research are identified along a polyrepresentation continuum.

2. The Principle of Polyrepresentation – A Holistic Approach to IR

The principle of polyrepresentation originates in work on establishing a cognitive theory for interactive IR (Ingwersen, 1992). It can be regarded as a result of an effort to demonstrate the *applicability* of this theory (Borlund, 2000). The principle of polyrepresentation was developed during the 1990s. The principle is fully expanded in the *Journal of Documentation* article from 1996, which remains the main publication on the principle (Ingwersen, 1996). Prior to that, the idea of polyrepresentation is mentioned throughout Ingwersen's book *Information Retrieval Interaction* from 1992 as a high precision tool, and an early version was presented at a SIGIR conference (Ingwersen, 1994).

A recent update on the principle appeared at the CoLIS4 conference (Ingwersen, 2002). This section briefly examines the cognitive foundation of the principle of polyrepresentation, outlines its main hypothesis, and gives examples of the kind of representations that can be captured from the cognitive space of the user and the information space of the IR systems respectively.

The cognitive foundation of the principle of polyrepresentation is very apparent in that it is assumed that every agent in IR contributes with their cognitive perceptions and interpretations of a given document as seen from their own context. Based on the cognitive viewpoint in Information Science (De Mey, 1980) such interpretations are viewed as the result of transformations of cognitive models or knowledge structures of the agents involved. When recorded they are regarded as representations of the documents. Obviously, the same document may be interpreted by many different agents, which can result in several representations of the same document, representations that have *different cognitive* origins.

For instance, an academic article (itself a representation of the author's knowledge structures at a given point in time) may be represented by controlled and uncontrolled indexing terms assigned by an indexer, as well as subsequently cited by various authors in different contexts. In addition, a number of representations may derive from the same agent but be of different *functional* nature, e.g., the article title, abstract, table captions, references etc. all generated by the author of a text. These two types of representations are normally associated with subject access to the content or aboutness of the documents. Supplementary representations may be generated based on the actions of *selectors* in relation to the isness of the documents (Ingwersen, 2002). These selectors are agents that are responsible for the availability and accessibility of documents, e.g., journal boards, reviewers, and employers etc. who bestow cognitive authority by, e.g., allowing a paper to appear in a given journal.

The cognitive foundation is also apparent in the way the searchers' situations are considered. An information need is not seen as a static entity in its own right but rather as part of a causal structure in which a work or daily-life task to be solved plays an essential role. Based on an analysis of the user's mental structures in relation to the information need it is proposed to extract a number of representations related to the cognitive space of the user. This is discussed in some detail in Section 3 below.

We therefore define a representation as any tangible and recordable entity that has occurred as the result of a transformation of knowledge structures of a cognitive agent. As the number of representations with different cognitive and functional origins rises it is evident that quite a lot of uncertainties and inconsistencies will arise in the representation of documents and information needs. Seen from a cognitive viewpoint this is inescapable, but not necessarily a disadvantage when exploited constructively.

One may view all the communication processes in IR as consisting of interchanges that take place at the *sign level*, i.e., at a linguistic surface level (Smeaton, 1992). When humans are part of IR activities the communication between generators and recipients of information may in addition take place at a *cognitive level*. Thereby the knowledge structures of the human recipient could potentially be affected and modified, and the searcher thus receives information to help solve her work or daily-life task. Because we may only communicate via signs, the information sent by a generator will always be subjected to a cognitive "free fall", and has to be re-interpreted by a human recipient to achieve communication at the cognitive level (Ingwersen, 1996, p. 6). This essential act of interpretation means that uncertainties and unpredictabilities become inherent features of *any* representation in IR. Much of the research and development work, e.g., in Library and Information Science, has been done to reduce such uncertainties and unpredictabilities. Controlled vocabularies like thesauri or extensive cataloguing rules are constructed to create more uniform document representations, as well as to allow users to reduce their uncertainties when formulating their needs.

The principle of polyrepresentation takes a holistic cognitive view by focussing simultaneously on the cognitive structures, and the representations that may be generated from these, both in the cognitive space of the user *and* in the information space of the IR systems. The principle of polyrepresentation is based on the following hypothesis:

> As all tangible representations in IR are the result of interpretations by the involved cognitive actors inconsistency and uncertainty are inherent in the representations. Therefore, if representations of different cognitive and functional origin point to a set of information objects in a particular seeking context, this cognitive overlap presents more relevant information objects than each independent representation. Further, the more different the cognitive representations producing an overlap are in time and by

cognitive or functional type, the higher the probability that such information objects are relevant. (Based on Ingwersen, 1996)

The principle of polyrepresentation thus represents an attempt to view the uncertainties and unpredictability as favourable to IR, and to exploit these actively rather than attempting to eliminate them. Inspired by, e.g., Sparck Jones (1990), polyrepresentation entails working through *intentional redundancy*, that is, to represent documents (or information needs) in multiple, complementary ways. We call this redundancy "intentional", since general and non-estimated redundancy in relation to representations of documents may not always be productive in IR. Rather the degree, mode and function of redundancy should be determined by knowledge of the current user's information behaviour. This means that special emphasis should be placed on documents in searcher-dependent overlaps created by intersecting representations of different cognitive and functional origins – cognitive overlaps (See also Figure 2 below).

The purpose of the principle of polyrepresentation is thus to facilitate the exploitation of a multitude of both cognitive and functional representations with focus on exploiting *different* functional representations from the same agent, as well as on combining representations from *different* cognitive agents in a structured framework. The following two sections examine the cognitive structures available in the cognitive space of the user and the information space of the system.

3. Polyrepresentation of the User's Cognitive Space

Early investigations into the formation of the information need have inspired the application of the principle of polyrepresentation on the cognitive space of the searcher. These investigations clearly demonstrate that there are different underlying cognitive reasons behind the development of an information need. This need is seen as the result of communication, sensing or thinking processes, which result in the realisation that something is missing for the solving of a problematic situation. This conception is inspired by the work of, e.g., Mackey (1960), Belkin (1980), Belkin, Oddy and Brooks (1982), and Dervin and Nilan (1986).

Four such cognitive structures influencing the user are shown in a bottom-up order of causality on the right hand side of Figure 1 below: a work (or daily-life) task/interest; a current cognitive state; a problem space, including a state of uncertainty; and an information need (Ingwersen, 1996).

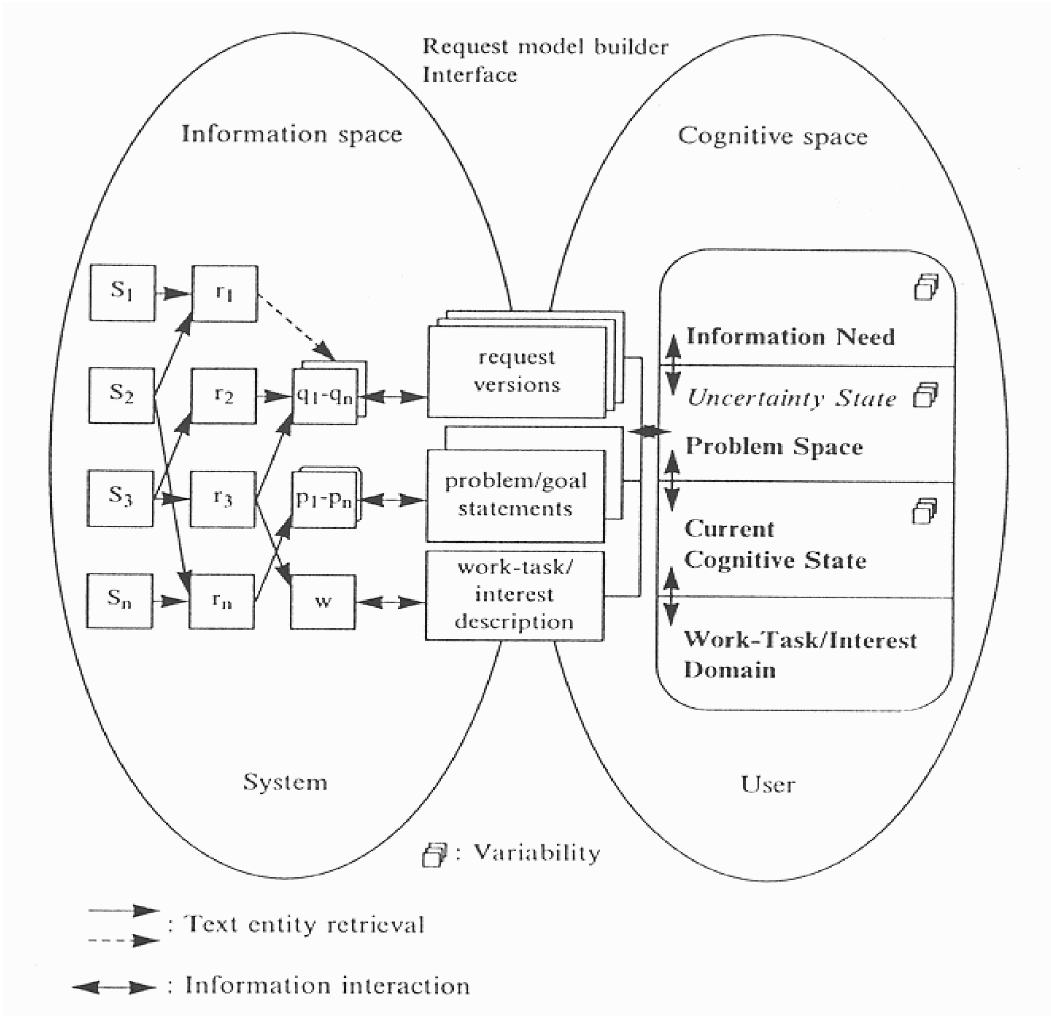

Figure 1. The Global Model of Polyrepresentation. (From Ingwersen, 1996, Figure 8, p. 37)

The four cognitive structures are the result of an analysis of the stability of the user's cognitive-emotional structures in relation to the information need, and places that need in a causal structure: The work task/interest perception, which is influenced by a common, professional or scientific domain and dominated by the individual's intentionality and cognitive-emotional state, causes a problematic situation, and, if this cannot be solved without external information, an information need. The current cognitive state includes the (limited) awareness or sense of what is desired. The current cognitive state, the problem space and the information need are all *variable*.

These three structures are easily affected by external input and/or thinking processes, while the work task/interest perception is set in a social context and may be more stable. For example, it may be related to a person's actual work situation as an engineer, to a person's continuing interest in certain aspects of Tolkien's artificial languages or other daily-life situations. The problem space corresponds to Belkin's ASK (Belkin, 1980; 1982) and is separated from the information need because the same problematic situation may give rise to several different information needs.

Based on earlier research it can be inferred that such cognitive-emotional structures have been manifested, and that some of them appear in quite well-defined forms (Ingwersen, 1996). However, as the four cognitive structures reside in the cognitive space of the user we do not have direct access to them during IR, and we have to work with *tangible representations* of them. Such representations or evidence could be extracted by human intermediaries or by interface functionalities during information interaction (indicated as a request model builder (RMB) interface in middle of Figure 1).

Polyrepresentation of the user's cognitive space may thus be achieved by capturing a number of different functional representations from the information seeker, as indicated in the middle of Figure 1. In an ideal situation, potentially at least three *functionally different* representations may be extracted at any one point in time (Ingwersen, 1996):

1. a 'what', i.e., a request version which includes what is currently known about the unknown (the wish or desire for information);
2. the 'why', i.e., a problem statement as well as
3. a work task or daily-life interest and domain description.

Because the underlying cognitive structures are variable over a session, different versions of each representation may occur over time (this temporal variability is indicated in Figure 1 as an increasing number of 'boxes'). Some of the captured representations may often appear to be similar, e.g., the problem statement and the work task description. This is a consequence of the fact that information needs may be well or ill defined, as well as more or less stable.

Evidence of emotional aspects, e.g., of uncertainty, doubt, satisfaction or relief, may be included in statements (Kuhlthau, 1991; 1993). These different types of information needs and their development are clearly demonstrated by the empirical studies (Belkin, 1984; Ingwersen, 1982), as is the role of the librarian in helping the user to define and refine her need. Such a set of representations extracted from the seeker's cognitive space provides a more fertile and *operational context* than a request version alone. The intention in the principle of polyrepresentation is that this enriched set of representations should be combined with each other and used as search terms during interactive IR, in order to achieve polyrepresentation of the information space.

4. Polyrepresentation of Documents in the Information Space of IR Systems

The information space of IR systems consists of two major components: the documents and the IT components that give access to the documents. The possible representations one may generate from each component are considered below. The *documents* are influenced by several different cognitive agents. Figure 2 below shows an overview of possible cognitively and functionally different representations of the documents. In addition, the figure may be seen as illustrating the sets of documents, and various cognitive overlaps between them that may be retrieved in relation to a request using each of the representations (Ingwersen, 2002).

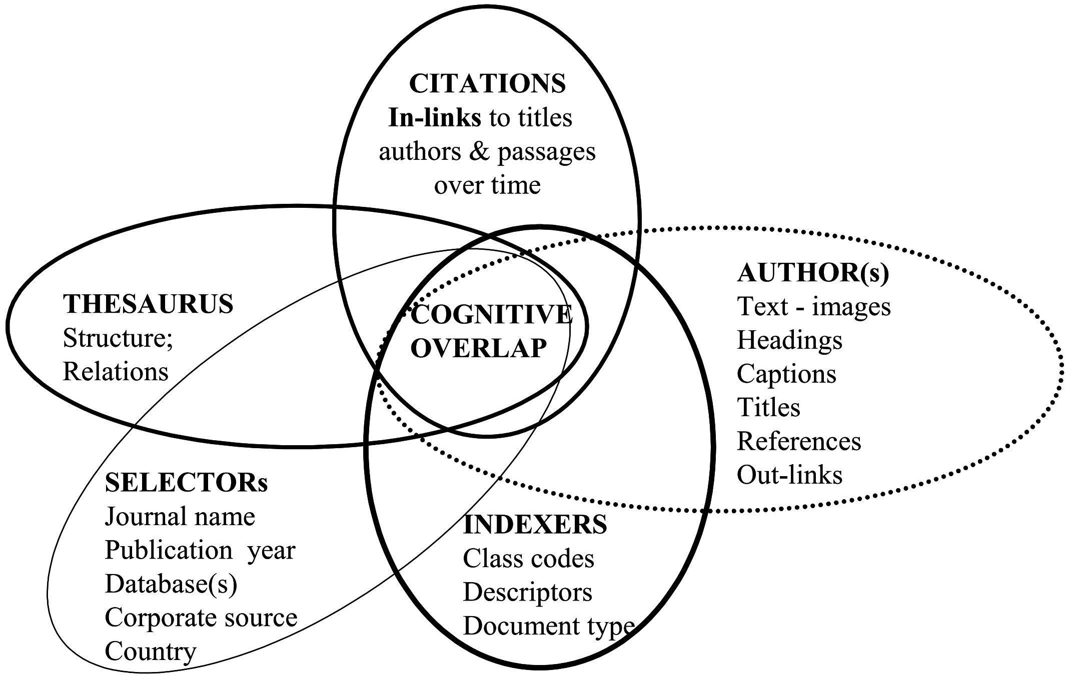

Figure 2. Polyrepresentative Cognitive Overlaps Generated from Cognitively and Functionally Different Representations of Documents. Retrieved Sets are Generated by One Retrieval Engine and Associated with One Searcher Statement (From Ingwersen, 2002, Figure 1, p. 294)

The contents of the documents reflect the cognitive structures of the *author(s)* in the form of signs, i.e., the transformations of the interpretations, ideas, and cognitive-emotional structures of the authors(s) with respect to their goals and intentionality. Depending on domain, media, and style, a range of functional representations can be identified in a single document. The representations in Figure 2 are related to academic full text documents – a particularly rich source for generating representations and hence ideal for experiments involving the principle of polyrepresentation.

The functional representations originating from the author that may be captured from the full text of the documents are shown on the right side of the figure. Many representations with strong functional characteristics are available because of the rhetorical structure of the academic documents, commonly organised in specific ways according to convention, e.g., introduction, theory, or methodological sections, results, discussion, and/or conclusions. Like presentation style, the *structural organisation* is domain and media-dependent and very useful as a *supplement* to subject matter. Aside from the structure of the documents, the section titles at different levels and the table and figure captions are examples of functionally different ways of representing a document. These have previously been applied for document representation (Wormell, 1981).

In addition, the *references* in the bibliographies may also be extracted from full text academic documents. The selection of particular references in a given document is seen as highly reflective of the situational factors that affect the author and her current cognitive state at the time of authorship. Similarly, they may signal a kind of situational appropriateness to a potential user who, in her particular situation, might agree to (parts of) the selection of references and thus find the document relevant. An additional opportunity, offered by academic documents in full text, is to identify the text surrounding the location in ci<u>ting</u> documents (if any) where the ci<u>ted</u> document is mentioned, and to use this as a representation of the cited document (See the top of Figure 2). The use in web search engines of the anchor text of hyperlinks on web pages as a representation of that web page, which receives the link, exploits the same idea (Brin and Page, 1998).

Aside from the author, another major cognitive agent who typically produces representations of academic documents is the *human indexers.* Their cognitive structures are represented by the index terms added to the original documents. Essentially they are the result of an intellectual interpretation of an author's text or images, often guided by predefined rules. Indexers typically select class codes and descriptors. These are taken from controlled vocabularies, e.g., a thesaurus (for which other human beings are responsible). Again, a thesaurus is an interpretation of the vocabulary and semantic relations of concepts within the domain covered by the thesaurus.

Consequently, the thesaurus constructor restricts the interpretation of the indexer and has an indirect influence on the representations made by the indexer. Apart from the indexer representations shown in Figure 2, uncontrolled terms and phrases (often called identifiers) may also be assigned. The identifiers represent the indexer's cognitive structures more directly, because the indexer chooses them freely. The thesaurus, the left-hand side, Figure 2, may in addition serve as a device for providing lead-in search keys (Bates, 1986), and as an automatic search key expansion tool.

A final group of representations that are indirectly influenced by a variety of cognitive agents are the *selectors* mentioned above. They are shown in the lower left corner of Figure 2, and contain structures that are *selective* and different from those of indexers, authors and users. Instead of aboutness, such features reflect *isness* – that a document actually exists. Most of the common bibliographic data or metadata thus belong to representations of isness by making available non-topical features connected to documents, depending on media, domain, and presentation style. They are the result of selection or assessment processes performed by various agents on documents and their authors over time (Ingwersen, 2002).

Examples of these include journal editors or a conference committee responsible for items that are selected for publication in a journal (or conference). The editor and not entirely the author determines the publication year or date. Such agents are affected by their social/academic context over time. The authors' affiliations also possess selective power, e.g., by hiring particular researchers.

5. Polyrepresentation of the IT Component of IR Systems

The IT components in the information space consist of representation techniques and indexing rules, matching algorithms, database architectures, search languages and computational logics produced by systems designers and producers. That is, the IT components are the main interest of researchers in the mainstream system-driven IR tradition. A given *retrieval engine* is an implementation of a particular combination of these. From a cognitive viewpoint a retrieval engine is the embodiment of the ideas of its creators. Thus cognitive overlaps between *different* retrieval engines can be studied. A large volume of research within the system-driven tradition clearly shows that the various best match IR techniques retrieve different but overlapping results, and the more alike the retrieval algorithms, the larger the overlap.

Among others Rajashekar and Croft (1995) investigated combining results from multiple algorithmic and operational index representations and retrieval strategies (engines) by application of the plausible inference network model of IR. They found indications of performance improvements in some combinations. The ranking produced by one best match IR system is ultimately a cognitive representation of the knowledge structures of its designers, and hence of interest to the principle of polyrepresentation: The simultaneous application of *several* different engines is consequently assumed to provide overlaps of objects of superior value to searchers than each single engine alone. The idea of data fusion in IR, as explored, e.g., by Belkin, Kantor, Fox and Shaw (1995), thus corresponds to aspects of polyrepresentation.

When the principle of polyrepresentation is applied for IR purposes, the various representations of the documents are matched with the representations of the user's information need by means of a retrieval engine. Figure 1 above illustrates the enriched platform with a range of possibilities for this match created by the principle of polyrepresentation. On the left hand side of the figure the objects containing information are shown as semantic entities (denoted by $S_1 \ldots S_n$), which may be whole documents or parts of documents[1]. Each of these may give rise to a number of representations with various cognitive and functional origins as discussed above ($r_1 \ldots r_n$) that may be matched with the functional representations of the user's information need, either the work task description (w), problems statements (p_1-p_n) or requests (q_1-q_n).

The w, ps and qs are in a form suitable for manipulation by the retrieval engine used, and there may be several versions of each as the cognitive space of the user changes over time. Note that Figure 1 demonstrates this for *one* retrieval engine only. As mentioned above different engines, e.g., one based on the vector space model (Salton, 1989) and one on the inference network model (Turtle and Croft, 1990), will retrieve and rank different documents, and the cognitive overlap between these may also be worth examination.

6. Problems and Potentials

[1] The left hand side of the figure is inspired by the inference network model (Turtle and Croft, 1990; Turtle, 1991)

Compared with the mainstream IR research tradition and the research carried out in the information seeking community, the principle of polyrepresentation offers a much broader approach than either two. The mainstream system-oriented IR research tradition focuses on document and request representation and matching algorithms, but not on the actual users of the system (Baeza-Yates and Ribeiro-Neto, 1999). Similarly, the user-oriented information seeking community focuses on the user's situation and seeking behaviour, but rarely on the IR systems involved. The principle of polyrepresentation stresses the importance of all agents and the interplay between them as a condition for achieving successful and optimal IR.

The potential of the principle is therefore to serve as a common theoretical framework for research that integrates the information seeking perspective on the users with the mainstream IR focus on designing and testing better IR systems. In addition, the validity of the assumptions expressed in the hypothesis above in Section 2 is clearly demonstrated by available research from several sub-disciplines in information science. See Ingwersen (1996, 2002) for a variety of examples from IR, information seeking, bibliometrics and scientometrics. The principle is not only a theoretical contribution, it also points to areas in which it can be applied and to research methodologies to be employed. This section examines the potentials and problems of putting the principle into practice, and points to challenges to be faced in the near future.

7. Studies of the Polyrepresentative Nature of IR

Relatively few studies have been based explicitly on the principle of polyrepresentation. Also large-scale interactive IR experiments with polyrepresentation implemented both on the system and user side are still lacking. One important reason for this may simply be that the principle, because of its holistic and comprehensive nature, requires a very large research effort to implement in its entirety. This is in part due to the fact that so far there has been a lack of test environments that are suitable for working with all the elements in the principle. Methodologically the principle of polyrepresentation is complicated to work with because at least two to three variables and their interplay are studied at the same time, e.g., {retrieval engine, user, user's context} or {retrieval engine, interface, user} (Ingwersen and Järvelin, 2004).

For instance, the main tracks in Text Retrieval Conferences (TREC) are based on the laboratory model, where users are not involved interactively (See, e.g., Harman, 1993). The TREC test corpora are large, but consist mainly of news articles with very limited possibilities for polyrepresentation of the documents. The interactive track at TREC has demonstrated, however, that users can form part of IR experiments in controlled settings (Hersh and Over, 2001). Although polyrepresentation of the documents is hard to achieve with the current document collections in TREC, there is no reason why it should be impossible to experiment with polyrepresentation on the user side. Information seeking research, on the other hand, is based on studying the user, and involves to an extent, e.g., user interaction with operational online systems. The document surrogates in these systems offer some possibilities for polyrepresentation of the documents. Research from this perspective would therefore at first glance seem better suited for experiments with the principle of polyrepresentation.

The main difficulty here is for end users to juggle all the possible representations and overlaps manually. Selecting which representations to combine and carrying out the necessary and complex Boolean combinations may lead to the cognitive overload – or simply prohibit searching. Automated interface functionalities, such as the request model builder (RMB) indicated in Figure 1 above, would thus be important for an operational system based on the principle of polyrepresentation. The RMB would question the user to extract several versions of the information need and the underlying work task, and assist in the matching of these with representations of the documents and prioritising them over the search session. These tools have, however, yet to be developed.

In the initial proposals of the principle of polyrepresentation (Ingwersen, 1994, 1996), the RMB was inspired by elaborate user modelling approaches such as those put forward in the MONSTART and MEDIATOR models (Belkin et al., 1987; Ingwersen, 1992). Very complex RMBs can be constructed from such models, but we propose to begin more modestly with simpler versions. The investigation of principles for how a RMB should function is central for future research on the principle of polyrepresentation. One initial approach to test would be to attempt to acquire different descriptions of the 'what', the 'why', and the perceived work task from users over search sessions.

A simple interface could interview users, e.g., first asking for what is needed (like the requests submitted to the present search engines), then enquiring why the information is needed (i.e., a description of the current problem to be solved), and finally asking for a description of the broader work task from which the current problem has arisen. These descriptions could then be matched against document representations according to general models of information seeking behaviour. As indicated in Figure 1 above, the description of the perceived work task/interest is likely to remain stable (at least for the duration of a search session), whereas the current problem might change over the session, giving rise to a number of different requests.

These kinds of assumptions could be part of the matching model. Even if no interactive interface is available to match the extracted descriptions in a polyrepresentative manner, such descriptions could be used to *simulate* an interactive system. Initial simulation studies could test which methods would be appropriate for matching different representations of the information need with document representations. This is similar to the simulation approach by White, Jose, Rijsbergen and Ruthven (2004) discussed below. An alternative source for such descriptions might be existing test collections:

From 2004 the IR topics of the INEX test collection[2] includes descriptions that are similar to the ones discussed above. The INEX 2004 topics are based on the idea of simulated work tasks as put forward by Borlund (2000, 2003): In order to be used in the interactive track of INEX, it was required that all INEX 2004 topics not only detail *what* is being sought for, but also *why* this is wanted, and in what *context* the information need has arisen. Simulated polyrepresentation studies can be carried out with static topics like these, although much of the realism would be lost due to the missing interaction. The results could inform us which features are important to include in an interactive RMB, and result in better prototypes to be used with real users.

More advanced RMBs could include the approaches to incorporating captured user behaviour into adaptive relevance feedback algorithms studied by Ruthven, Lalmas and van Rijsbergen (2003) and White, Jose, Rijsbergen and Ruthven (2004). Ruthven, Lalmas and van Rijsbergen (2003) used a novel approach to

---

[2] Every year from 2002 the Initiative for the Evaluation of XML Retrieval (INEX) constructs an IR test collection from 12,107 full text articles from the IEEE Computer Society's 20 journals. For more information see Gövert and Kazai (2003) and the INEX web site http://inex.is.informatik.uni-duisburg.de/2005/

explicit relevance feedback: information on *how relevant* the user finds a retrieved document is combined with *how old* the assessments are, giving priority to the most recent assessments. The explicit assessments are used to select terms for query expansion and adapts to the user's behaviour over the search session. White, Jose, Rijsbergen and Ruthven (2004) studied how *implicit feedback* detected from browsing behaviour can be used as input to relevance feedback. The parts of documents the users choose to display and the order of these are part of the implicit feedback. They are used to adapt the query to the user's information need. Both studies demonstrate how better representations of the user's current information need can be obtained through interaction, and that a number of representations can be extracted over a session. These representations can be combined with the three different types of descriptions discussed above, in particular with the perceived work task description. Alternatively, the latest version can be used directly as an up-to-date representation of the request and/or the current problem to be solved.

8. The Matching Problems in Creating Cognitive Overlaps

Even with a working RMB prototype that could interview the user and extract different representations of her cognitive space, the question of how to match representations and generate cognitive overlaps may not be straightforward. The hypothesis above in Section 2 offers some advice, but only of a general nature (i.e., to match the ones that are most different in cognitive or functional type –

and probably also temporally different). We regard this issue as one to be solved largely by empirical research. Although the principle of polyrepresentation has not resulted in a large body of empirical research that deal with all its elements, a few recent studies have worked with some of them. Below, four such studies are examined. All four have implemented some form of polyrepresentation of the documents, but none have attempted to represent the user's cognitive space in several ways – only a single, static version of the information need was used. Two of the studies were carried out in operational online databases (Christoffersen, 2004; Larsen, 2002), while the other two used test collections (Larsen, 2004; Skov, Pedersen, Larsen and Ingwersen, 2004).

Larsen (2002) proposes a new strategy for searching via citations, the so-called 'boomerang effect'. The strategy was tested in a small experiment carried out in the online version of the Science Citation Index (SCI). One test person provided three information needs based on his current research interests as a researcher at a hospital. The retrieved documents were subsequently assessed for their usefulness in relation to the needs. As the SCI does not contain the full text of the documents the number of representations used was limited: titles and abstracts by the author, and Keywords Plus (automatically assigned identifiers) and the network of references and citations by other cognitive agents.

The experiment was Boolean because of the online setting, and only static versions of the information needs were used. Larsen used the principle of polyrepresentation as inspiration for selecting representations, for the automatic identification of seed documents for the citation search, and for refining the results of the citation search strategy. This is an extension of Pao's (1994) investigations.

In brief, the strategy extracted the references from document sets retrieved from several different representations, and identified seed documents from the cognitive overlap between these. The seed documents were then used as starting point for a forward chaining through the citation network to identify documents that refer to the seed documents. The results of the boomerang effect consisted of a number of retrieved document sets ordered in a polyrepresentative overlap structure. They showed that the overlaps generated by many representations consistently had higher precision.

Although it was possible to implement the strategy for the purpose of the experiment, Larsen notes that with more than three or four representations the number of overlaps and the effort required to handle them increase dramatically. He also experienced problems with the initial query formulations, which had to be expanded in order to fit the individual representations. The expansion was also necessary in order to ensure that the sets were sufficiently large to produce an overlap in the first place. Finally, the output was a semi-ranked list of document sets with no internal ranking. This can present problems to users if a set is large, and this kind of output makes it very hard to compare the strategy to other IR approaches, especially best match systems.

Christoffersen (2004) used the online versions of Medline, Embase and SCI to test the proportions of relevant documents in the overlaps between the three databases. The representations had strong cognitive differences: Title/abstract words (from authors extracted from Embase) vs. MeSH terms (from indexers extracted from Medline) vs. searching by citations (from citing authors in SCI). The relevance assessments were by subject experts, and the results showed that the degree of overlap (i.e., the number of sets a document appeared in) correlated strongly with the percentage of relevant items in a set. Again the study was Boolean because of the online setting, but as only three representations were used no serious problems were experienced in handling the overlaps. The intersections involved did, however, reduce the number of documents in the overlaps to less than 14 % of the total number of documents retrieved. This demonstrates polyrepresentation as a strong precision tool.

Both studies were on a small scale and used only a few representations, and the promising results of both are therefore remarkable. This may be interpreted as a consequence of the strong cognitive differences between the representations used. An equally important factor is the size of the database involved: in both cases the systems were operational and contained several million records each. The principle of polyrepresentation could therefore reduce the uncertainty associated with each individual representation, and still create overlaps that were not empty.

Skov, Pedersen, Larsen and Ingwersen (2004) set out to test elements of the principle of polyrepresentation in a best match setting. A small test collection containing 1239 Medline records augmented with references and citations, but without the full text of the documents, was used (Shaw, Wood and Tibbo, 1991). Despite its small size the collection offered several cognitive and functional representations: words from titles and abstracts (from the author), Minor and Major MeSH headings (by indexers) as well as references and in-going citations (by citing authors).

Skov et al. selected four of the representations and generated all possible cognitive overlaps between them to study the performance of each one in their experiment. Two types of queries were tested in a best match system: natural language queries and highly structured queries. Both types used Boolean operators to identify overlaps. The highly structured queries also contained indications of query facets and phrases, and had synonyms added intellectually from MeSH. These additions were the results of Skov et al’s efforts to, on the one hand, improve the quality of the document sets, and on the other, ensure that the overlaps were non-empty. Results showed that overlaps generated by several representations had higher precision than those generated from few representations for both query types.

Marked differences were also found between representations; in particular the results indicate an increase in precision when documents identified by a citation search strategy formed part of an overlap, stressing the importance of using representations that have strong interpretative differences. In all cases the highly structured queries achieved higher precision than the natural language queries, which is explained as a consequence of generating overlaps in a best match system: because the natural language queries only require one search term from the query to be present, the retrieved sets of documents, and thus the overlaps, may contain documents with very little relation to the information need. In contrast to this, the highly structured queries ensure that all facets of the queries are present in each cognitive overlap. Skov, Pedersen, Larsen and Ingwersen's (2004) results indicate that the quality of the initial sets from which overlaps are created can be improved and better results achieved with the principle of polyrepresentation, but only after extensive work on refining and expanding the queries. This can also ensure that recall does not suffer too much.

Larsen (2004) tested a best match version of the boomerang effect using the INEX computer science test collection. Because of the complex full text XML structure a number of functional representations could be extracted from the documents: title, abstract, author keywords, cited titles (from the reference list) as well as figure and table captions and the introduction and conclusion sections. In addition, the documents were represented by descriptors from the INSPEC thesaurus and uncontrolled identifiers assigned by INSPEC indexers.

The boomerang effect was tested against two baselines: a bag-of-words index where all the representations were mixed into one, and a polyrepresentation baseline, which gave higher weights to documents retrieved from several representations and required that documents were retrieved from at least two representations. The same unstructured queries were used in all three runs. Results showed that the bag-of-words baseline out-performed the other two, and that the polyrepresentation baseline performed slightly better than the citation search strategy in the boomerang effect.

Strict Boolean overlaps were not enforced in any of the strategies; Larsen had instead chosen to rely on thresholds to limit the size of the sets. It should be noted that the best performance of the latter two strategies were obtained at relatively low thresholds, i.e., when the sets retrieved from each representation contained few documents. This may be explained similarly to Skov et al's results: as the best match system only requires at least one of the query terms to be present in the retrieved documents, only the documents at the top of the rank have a sufficiently strong relation to the information need.

9. Cognitive Overlaps in Best Match IR Systems

The approaches in Skov, Pedersen, Larsen and Ingwersen (2004) and Larsen (2004) both produced ranked output, which has the advantage that it can be compared to standard best match IR methods. Best match methods are not, however, unproblematic when trying to generate strong cognitive overlaps. Best match systems will most often place the documents that contain all the query keys at the top of the ranked retrieval output, but will also include any document that contains just one of the query keys at lower positions of the rank. In addition, if a query key occurs very rarely in the database, but very frequently in a particular document, this document will be placed in the top of the rank because of the *tf*idf* weighting scheme, regardless of whether it contains any of the other query keys. The combination of partial match and ranked output is one of the main advantages of best match systems over exact match systems (Belkin and Croft, 1987).

However, in relation to the creation of overlaps in the principle of polyrepresentation there is a risk that the quality of the sets that the cognitive retrieval overlaps are based on as a whole is too low. For instance, with two search keys there is the risk that only the first of them is retrieved by some of the lower ranking documents from one representation, and the second in the lower ranks from another representation. Therefore proper polyrepresentation in the true sense of the concept cannot be achieved.

The results by Skov, Pedersen, Larsen and Ingwersen (2004) and Larsen (2004) demonstrate that unstructured applications of the principle of polyrepresentation are not likely to result in performance improvements – rather a decrease in performance can be expected, at least when simplistic fusion strategies such as those in Larsen (2004) are used. Thus, the implementation of the principle of polyrepresentation in best match systems is not straightforward. It seems that some structure like that offered by Boolean logic is needed to ensure the quality of the cognitive overlaps as seen in Skov, Pedersen, Larsen and Ingwersen (2004). The intention behind the principle of polyrepresentation is, however, that it should be applicable in both exact and best match settings. This calls for further investigations of which best match techniques can be fruitfully adapted to work with the principle.

One candidate would be the inference network model proposed by (Turtle, 1991; Turtle and Croft, 1990). The basic idea in the model is to view the retrieval process "…as an evidential reasoning process in which multiple sources of evidence about document and query content are combined to estimate the probability that a given document matches a query" (Turtle and Croft, 1992, p. 280). One advantage of this model is that it can represent many IR approaches and combine them in a single framework. Such a framework will allow experiments that examine the effect of generating cognitive overlaps with a variety of matching methods, both exact and best match approaches and their combinations.

The aim will be to find matching methods that will allow ranked output and, at the same time, ensure the quality of the cognitive overlaps. Another line of research to follow is investigations similar to Skov et al., where all possible overlaps between a range of different representations are examined to identify the most promising combinations. This also includes investigations of which representations do *not* intersect. For instance, Skov et al. decided to fuse the titles and abstracts into one representation because they tended to retrieve the same documents in their experiment. This becomes even more important when richer document corpora are used, e.g., the INEX corpus where a large range of functional and cognitive representations can be generated from the documents.

As shown by Skov, Pedersen, Larsen and Ingwersen (2004) it may also be important to attempt to improve the quality of the document sets before the intersection to generate a cognitive overlap. Therefore different query expansion techniques are an important and necessary issue in research on polyrepresentation. It is important to note that the different representations may require different kinds of expansion in order to be successful.

10. The Polyrepresentation Continuum

As shown by the studies discussed above the matching of representations with different cognitive and functional origins of both document and the user's information need is a very essential research issue for the principle of polyrepresentation. The main challenge is to identify methods of achieving a flexible match of representations while still retaining some of the power and quality of the Boolean logics. The studies analysed above show that this is not straightforward, and that much more work needs to be done on the matching of representations before the potentials of the principle can be fully realised. Rather than using either exact match or best match approaches it is our belief that a combination of methods is needed.

Therefore we propose the idea of a *polyrepresentation continuum* as illustrated in Figure 3 below. The continuum is useful as a model for discussing how structured a given implementation of the principle of polyrepresentation is, and may guide the direction of further work on the principle.

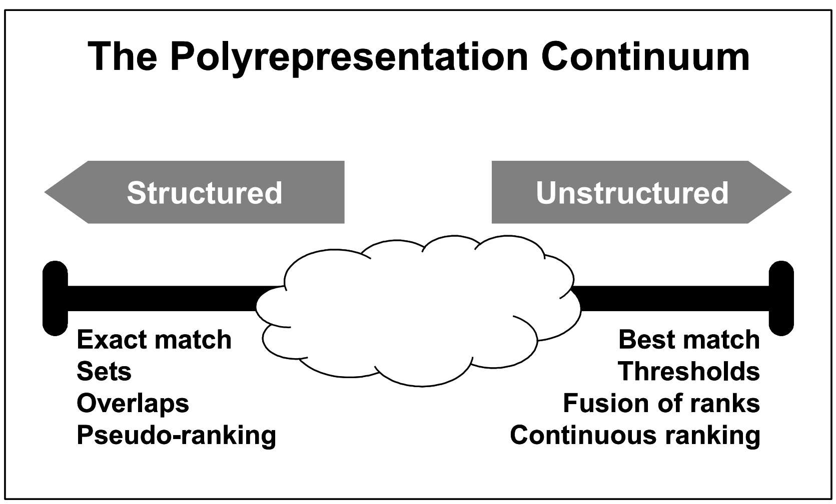

Figure 3. The Polyrepresentation Continuum (Larsen, 2004)

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 retrieved documents as input for polyrepresentation. Rather than generating overlaps between sets, the implementations at the unstructured end of the polyrepresentation continuum will fuse the ranks to produce a final ranked output, perhaps aided by thresholds to provide the necessary quality. Between the two poles there is a continuum of polyrepresentative constructs going from highly structured implementations to highly unstructured implementations.

The implementations in Larsen (2002) and Christoffersen (2004) are placed at the structured pole of the continuum. The polyrepresentation baseline in Larsen (2004) is placed at the unstructured end. The cloud in the middle of the continuum illustrates the current status where we do not know how to match the representations in a flexible and effective manner, and identifies the main challenge for future research on the principle of polyrepresentation: to identify flexible matching methods that can generate high quality cognitive overlaps from a variety of representations.

Skov, Pedersen, Larsen and Ingwersen (2004) represents a constructive attempt to move from an exact match approach towards the unstructured pole of the continuum. Their highly structured queries that are run in a best match system can be seen as a second dimension of the continuum. Further moves toward the unstructured pole may include structured query types as investigated in best match systems by Kekäläinen and Järvelin (1998), where facets of the information needs are identified and expanded using a thesaurus. Query expansion and query adaptation to individual representations seem important and might lead to more formal IR models which incorporate differentiated normalisation and weighting for different representations. It may for instance be that the characteristics of certain representations are underrepresented by the standard *tf*idf* weighting scheme. Instead of calculating *idf* values in relation to the whole document, they might be calculated in relation to each representation in order to capture their individual characteristics more successfully.

In total we may observe at least three dimensions of the polyrepresentation continuum: 1) the degree of structure in the polyrepresentation matching modus, Figure 3; 2) the degree and nature of structure within queries by application of facets; and 3) the modus of expansion or modification of queries, like expansion of facets. How influential the latter two dimensions are on the entire range of matching principles is yet to be investigated in a detailed manner.

11. Conclusion

The principle of polyrepresentation is a coherent and comprehensive cognitive framework that can be applied simultaneously to the cognitive space of the user and the information space of IR systems. The principle has the potential to guide the design of interactive IR systems that take full advantage of the available document representations *and* user's context to improve retrieval performance.

However, before this can be achieved a number of issues need further investigation. Among these are simulation studies that test which methods would be appropriate for matching different representations of the user's cognitive space with document representations. Such simulations could apply simulated work task situations; or they could explore exhaustively the possibilities of a number of controlled variables and thus simulate all achievable combinations. Investigations involving test persons and experimental laboratory tests (simulations) must take into account the dependency of domains, media and representation styles.

Studies of the principles for how a Request Model Builder should function would be fruitful, and how to match representations to generate strong cognitive overlaps, especially in best match settings. The latter issue is illustrated by the polyrepresentation continuum. It points to the investigation of flexible and powerful hybrid matching of representations as a challenge and opportunity for future research along these lines.

12. References

Baeza-Yates, R. and Ribeiro-Neto, B. (1999). *Modern Information Retrieval*. Harlow: Addison Wesley. xvi, 513 p. ISBN: 0-201-39829-X.

Bates, M. J. (1986). Subject Access in Online Catalogs: A Design Model. *Journal of the American Society for Information Science*, 37(6), 357-376.

Belkin, N. J. (1980). Anomalous States of Knowledge as Basis for Information retrieval. *The Canadian Journal of Information Science*, *5*, 133-143.

Belkin, N. J. (1984). Cognitive Models and Information Transfer. *Social Science Information Studies*, *4*, 111-129.

Belkin, N. J. Kantor, P. Fox, E. A. and Shaw, J. A. (1995). Combining the Evidence of Multiple Query Representations for Information Retrieval. *Information Processing and Management*, *31*(3), 431-448.

Belkin, N. J. Oddy, R. and Brooks, H. (1982). ASK for Information Retrieval: Part 1: Background and Theory. *Journal of Documentation*, *38*, 61-71.

Belkin, N. J. Borgman, C. L. Brooks, H. M. Bylander, T. Crofts, W. B. Daniels, P. Deerwester, S. Fox, E. A. Ingwersen, P. and Rada, R. (1987). Distributed Expert-Based Information Systems: An Interdisciplinary Approach. *Information Processing and Management*, 23(5), 395-409.

Belkin, N. J. and Croft, W. B. (1987): Retrieval Techniques. In: M. E. Williams (Ed.), *Annual Review of Information Science and Technology (ARIST), 22, 1987*. Amsterdam: Elsevier Science, p. 107-145.

Borlund, P. (2000): *Evaluation of Interactive Information Retrieval Systems*. Åbo: Åbo Akademi University Press. vi, 276 p. ISBN: 951-765-049-3. (PhD dissertation)

Borlund, P. (2003): The IIR Evaluation Model: A Framework for Evaluation of Interactive Information Retrieval Systems. *Information Research*, *8*(3), paper no. 152. [http://informationr.net/ir/8-3/paper152.html, visited 6-8-2003]

Brin, S. and Page, L. (1998). The Anatomy of a Large-Scale Hypertextual Web Search Engine. *Computer Networks and ISDN Systems*, *30*(1-7), 107-117. [http://www-db.stanford.edu/pub/papers/google.pdf (preprint), visited 16-5-2003]

Christoffersen, M. (2004). Identifying Core Documents with a Multiple Evidence Relevance Filter. *Scientometrics*, *61*(3), 385-394.

De Mey, M. (1980). The Relevance of the Cognitive Paradigm for Information Science. In: O. Harbo and L. Kajberg (Eds.), *Theory and Application of Information Research: Proceedings of the Second International Research Forum on Information Science, 3-6 August 1977, Royal School of Librarianship, Copenhagen*. London: Mansell, p. 49-61.

Dervin, B. and Nilan, M. (1986). Information Needs and Uses. In: M. E. Williams (Ed.), *Annual Review of Information Science and Technology (ARIST), 21, 1986*. New York: Knowledge Industry Publications, p. 3-33.

Gövert, N. and Kazai, G. (2003). Overview of the Initiative for the Evaluation of XML Retrieval (INEX) 2002. In: N. Fuhr, N. Gövert, G. Kazai and M.

Lalmas (Eds.), *Proceedings of the First Workshop of the Initiative for the Evaluation of XML Retrieval (INEX): December 9-11, 2002, Schloss Dagsthul, International Conference and Research Centre for Computer Science*. [Sophia Antipolis]: [ERCIM], p. 1-17. ([ERCIM Workshop Proceedings]) http://qmir.dcs.qmul.ac.uk/inex/Papers/final_overview_Goevert_etal.pdf, visited 27-7-2003]

Harman, D. (1993). Overview of the First TREC Conference. In: R. Korfhage, E. Rasmussen and P. Willett (Eds.), *SIGIR'93: Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. New York: ACM Press, p. 36-47.

Hersh, W. and Over, P. (2001). Interactivity at the Text Retrieval Conference (TREC). *Information Processing & Management*, 37(3), 365-367. (Editorial: Introduction to Special Issue About the Interactive Track at TREC)

Ingwersen, P. (1982). Search Procedures in the Library Analysed From the Cognitive Point of View. *Journal of Documentation*, *38*(3), 165-191.

Ingwersen, P. (1992): *Information Retrieval Interaction*. London: Taylor Graham. x, 246 p. ISBN: 0-947568-54-9. [http://www.db.dk/pi/iri/, visited 1-11-2003]

Ingwersen, P. (1994). Polyrepresentation of Information Needs and Semantic Entities: Elements of a Cognitive Theory for Information Retrieval Interaction. In: W. B. Croft and C. J. van Rijsbergen (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.

Ingwersen, P. (1996). Cognitive Perspectives of Information Retrieval Interaction: Elements of a Cognitive IR Theory. *Journal of Documentation*, *52*(1), 3-50.

Ingwersen, P. (2001). Cognitive Information Retrieval. In: M. E. Williams (Ed.), *Annual Review of Information Science and Technology (ARIST), 34, 1999-2000*. Medford, NJ: Information Today, vol. 34, p. 3-51.

Ingwersen, P. (2002). Cognitive Perspectives of Document Representation. In: H. Bruce, R. Fidel, P. Ingwersen and P. Vakkari (Eds.), *Emerging Frameworks and Methods: CoLIS4: Proceedings of the Fourth International Conference on Conceptions of Library and Information Science, Seattle, WA, USA, July 21-25, 2002*. Greenwood Village, Colo.: Libraries Unlimited, p. 285-300.

Ingwersen, P. and Järvelin, K. (2004). Context in Information Retrieval. In: T. J. D. Bothma and A. Kaniki (Eds.), *Prolissa 2004. Proceedings of the 3rd Biennial DISSAnet Conference, Pretoria, 28-29 October 2004*. Pretoria: Infuse, p. 301-310.

Kekäläinen, J. and Järvelin, K. (1998). The Impact of Query Structure and Query Expansion on Retrieval Performance. In: W. B. Croft, A. Moffat, C. J. van Rijsbergen and J. Zobel (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.

Kuhlthau, C. C. (1991): Inside the Search Process: Information Seeking From the User's Perspective. *Journal of the American Society for Information Science*, *42*(5), 361-371.

Kuhlthau, C. C. (1993). A Principle of Uncertainty for Information Seeking. *Journal of Documentation*, *49*(4), 339-355.

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*. Copenhagen: Royal School of Library and Information Science. XIII, 297 p. ISBN: 87-7415-275-0. (PhD dissertation) [http://www.db.dk/blar/dissertation, visited 11-2-2004]

Mackey, D. M. (1960). What Makes the Question? *The Listener*, *63*(May 5), 789-790.

Pao, M. L. (1994). Relevance Odds of Retrieval Overlaps From Seven Search Fields. *Information Processing and Management*, *30*(3), 305-314.

Rajashekar, T. B. and Croft, W. B. (1995). Combining Automatic and Manual Index Representations in Probabilistic Retrieval. *Journal of the American Society for Information Science*, *46*(4), 272-283.

Ruthven, I., Lalmas, M. and Rijsbergen, K. v. (2003). Incorporating User Search behavior into Relevance feedback. *Journal of the American Society for Information Science and Technology*, *54*(6), 529-549.

Salton, G. (1989): *Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer*. Reading, Mass.: Addison-Wesley. xiii, 530 p. ISBN: 0-201-12227-8.

Shaw, W. M. Jr., Wood, J. B. and Tibbo, H. R. (1991). The Cystic Fibrosis Database: Content and Research Opportunities. *Library & Information Science Research*, *13*, 347-366.

Skov, M. Pedersen, H. Larsen, B. and Ingwersen, P. (2004): Testing the Principle of Polyrepresentation. In: P. Ingwersen, C. J. van Rijsbergen and N. Belkin (Eds.), *ACM SIGIR 2004 Workshop on "Information Retrieval in Context"*. Sheffield: [University of Sheffield], p. 47-49. (Workshop proceedings) [http://ir.dcs.gla.ac.uk/context/, visited 7-2-2005]

Smeaton, A. F. (1992): Progress in the Application of Natural Language Processing to Information Retrieval Tasks. *Computer Journal*, *35(*3), 268-278. [http://www3.oup.co.uk/computer_journal/hdb/Volume_35/Issue_03/, visited 17-2-2005]

Sparck Jones, K. (1990). *Retrieving Information or Answering Questions*. London: The British Library Board. (British Library Annual Research Lecture ; 8)

Turtle, H. and Croft, W. B. (1990). Inference Networks for Document Retrieval. In: J.-L. Vidick, (Ed.), *Proceedings of the 13th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval: Brussels, Belgium, 5-7 September, 1990.* New York: The Association for Computing Machinery, p. 1-24.

Turtle, H. R. and Croft, W. B. (1992). A Comparison of Text Retrieval Models. *The Computer Journal*, *35*(3), 279-290. [http://www3.oup.co.uk/computer_journal/hdb/Volume_35/Issue_03/, visited 22-8-2003]

Turtle, H. R. (1991). *Inference Networks for Document Retrieval*. [Massachusetts]: University of Massachusetts. xiii, 198 p. (PhD dissertation) [http://ciir.cs.umass.edu/pubfiles/ir17.pdf, visited 4-6-2003]

White, R. W. Jose, J. M. Rijsbergen, C. J. and Ruthven, I. (2004). A Simulated Study of Implicit Feedback Models. In: S. McDonald and J. Tait (Eds.), *Advances in Information Retrieval, 26th European Conference on IR Research, ECIR 2004, Sunderland, UK, April 5-7, 2004, Proceedings*. Springer, p. 311-326. (Lecture Notes in Computer Science ; LNCS 2997) [http://www.cis.strath.ac.uk/~ir/papers/ecir_implicit.pdf (preprint)]

Wormell, I. (1981). SAP - A New Way to Produce Subject Descriptions of Books. *Journal of Information Science: Principles and Practice*, *3,* 39-43.