# References and citations in automatic indexing and retrieval systems – experiments with the boomerang effect

Birger Larsen

PhD thesis from Department of Information Studies
Royal School of Library and of Information Science, Denmark

# References and citations in automatic indexing and retrieval systems – experiments with the boomerang effect

Birger Larsen

PhD thesis from Department of Information Studies
Royal School of Library and of Information Science, Denmark

**CIP – Cataloguing in Publication**

**Larsen**, Birger
References and citations in automatic indexing and retrieval systems – experiments with the boomerang effect / Birger Larsen. – Copenhagen: Department of Information Studies, Royal School of Library and Information Science, 2004. xiii, 297 p.
Available: http://www.db.dk/blar/dissertation
ISBN 87-7415-275-0

ISBN 87-7415-275-0

# Referencer and citationer i automatiske indekserings- og genfindingssystemer – eksperimenter med boomerangeffekten

Birger Larsen

Ph.d.-afhandling fra Institut for Informationsstudier
Danmarks Biblioteksskole

# Acknowledgments

Even though this volume records my own work, many people have contributed to this dissertation, directly or indirectly, without whom it never could have been finished. For this I thank you all.

At conferences, seminars and visits at other institutions I have had the pleasure of discussing my work with many knowledgeable scholars and PhD students. Special thanks go to the FIRE research group and the Department of Information Studies at Tampere University, Finland where I was welcomed heartily by everyone during my external visit. Thanks to Professor Kalervo Järvelin for lengthy discussions, good ideas, and thorough criticism of my work, to Jaana Kekäläinen for being honest and constructive about the realism of the project from the very beginning and much direct and indirect help, to Heikki Keskustalo for a unique and immediate understanding of my practical problems, and to everyone else at the department for a good and stimulating time in and outside saunas. I have also received encouragement and interest in my work from Professors Irene Wormell, Olle Persson, and Erik Frøkjær, who offered much needed help at a critical time, as well as from Ronald Rousseau, Wolfgang Glänzel, and Mounia Lalmas. Many thanks also to the Nordic PhD students, past and present, with whom I have shared so many good discussions and experiences at PhD courses, especially to Susanna Käränen, Fredrik Åström Wirthig, Rikard Danell, and Kai Haltunen up there in the North, as well as Jesper Wiborg Schneider, Lennart Björneborn, Jack Andersen and Frank Sejer Christensen in Denmark.

I owe many thanks for the invaluable direct help from Henrik L. Jørgensen with relevance assessment for the pre-experiment, for Petteri Kangaslampi in Tampere with the creation of candidate INEX topics, for the Danish branch of Dialog for generous access to their files over the years, for Victor Lavrenko at UMASS for updated InQuery files, and for the *like* program source code kindly lent by the Department of Information Studies at Tampere University. I would also like to thank Norbert Fuhr and Mounia Lalmas for arranging INEX brilliantly and listening and acting in matters of relevance judgements, and Norbert Gövert for pointing to a much needed document corpus.

I am very grateful for the very intensive efforts of Kalervo Järvelin, Morten Hertzum and Pia Borlund as my opponents. Your thorough and reflective comments have increased my understanding of the dissertation a great deal.

I would also like to express my gratitude to our Head of Department Mona Madsen and our departmental secretary Susanne Acevedo for support and encouragement throughout the project, to librarians Allan Dideriksen, Jytte Halling, Charlotte Kristoffersen and the rest of the staff at the Library of the Royal School of Library and Information Science for prompt and very consist help, to the caretakers at the Royal School of Library and Information Science Ole Ekenberg, Bjarne Toft Nielsen, Anders Andersen, John Jørgensen, and Jannick Thygesen for much help in practical matters, and to the staff at the IT-service at the Royal School of Library and Information Science Morten Gemzøe, Pia Brade, John C. Hansen, and James Hindsgavl Brink for trying to help with many out of the ordinary requests. A special thank to past and present members of the TAPIR research group, Peter Ingwersen, Pia Borlund, Marianne Lykke Nielsen, Erik Thorlund Jepsen, Haakon Lund, Piet Seiden, Lennart Björneborn and Jesper Wiborg Schneider at the Royal School of Library and Information Science for good discussions, support, collaboration and for wanting to do something, and to our laboratory assistant Jacob Andresen for being consistently curious and enthusiastically creative through many, long days of trying to make the boomerang effect run on our machines. Also thanks to numerous masters students who have brought my thoughts a step further with critical questions, and especially to Mette Madsen (now Skov) and Henriette Pedersen, who, through very hard and persistent work, have moved the ideas further and shown a way forward.

A very special thank is due to Pia Borlund for always being there and for really teaching how to be a PhD student. You are a very treasured colleague and a true friend.

The biggest credit is due to my supervisor Professor Peter Ingwersen for trusting that I could do the PhD (and many other things), and for backing me up all the way without hesitating a single time. You have given me constant and frequent intellectual stimulation, and always taken time to consider my problems. I am grateful for you as a dedicated and enthusiastic supervisor, and for you as a cherished friend.

Last by not least, I would like to thank my friends and my family, in particular, my parents, my grandparents, my sister Sanne, and especially Katrine for patience and endurance, and for always being there and believing in me.

## Abstract

The objective of the present dissertation work is to investigate the use of references and citations as an integrated part of automatic indexing and retrieval techniques operating on scientific full text documents. There are two main motivations behind the dissertation: *Firstly*, the few scattered studies from both the system-driven and the user-oriented tradition that have investigated references and citations for information retrieval (IR) purposes, have generally shown promising results. *Secondly*, scientific documents are for the first time becoming available in large quantities in electronic form. This offers new possibilities for combining conventional automatic indexing and retrieval techniques with the exploitation of references and citations in IR.

Analytical and empirical investigations are carried out with the aim of investigating which factors that affect the behaviour and performance of automatic indexing and retrieval techniques given that references and citations are an integrated part of the document representation of scientific full text documents in the IR system. It is investigated *why* references and citations might be useful in IR in an analytical review of the literature on citer motivations and citation behaviour. A normative and a social constructivist position on the motives of authors in citing other work is identified. The most unifying theoretical explanation is found to Small's (1978) notion of references as 'concept symbols' that stand for an idea or concept that is being used in the course of an argument. This explanation is found attractive from the point of exploiting references and citations in IR, because regardless of whether or not some references are omitted, forgotten, biased etc. the references *actually given* function as symbols for a concept. This can explain why references and citations are useful in IR: Because the references represent concepts when used for document representation, they are well-suited for IR purposes as long as the user's information need can be expressed in the same concepts, i.e., as seed documents. A review of earlier studies which have employed references and citations showed that the most common use of them for IR purposes was as *seed documents* in a forward chaining, where documents that refer to the seed document are retrieved. The ability of these seed documents to express the user's information need was found to be the main factor affecting the behaviour and performance of references and citations in IR.

The method proposed in the dissertation – the so-called boomerang effect – represents an attempt to eliminate the need for the user to specify seed documents intellectually, and instead it identifies these automatically based on a natural language request. Inspired by the theory of polyrepresentation (Ingwersen, 1996) the boomerang effect extracts and prioritises seed documents from a range of different cognitive and functional representations of the documents, and retrieves other documents that cite these. The best match boomerang effect is placed at the un-structured pole of the polyrepresentation continuum that is proposed in the dissertation as an extension to the theory of polyrepresentation.

In addition, the boomerang effect serves as the framework for the empirical experiments, which was carried out as a system-driven laboratory experiment with the INEX test collection (Gövert and Kazai, 2003). The main experiment showed that it is indeed possible to retrieve relevant documents through the network of references and citations with automatic indexing and retrieval techniques. The boomerang effect performs as well as a polyrepresentation baseline without citations, but both displayed inferior performance compared to a conventional bag-of-words baseline. A number of factors were investigated to examine their influence on the behaviour and performance of the boomerang effect. It was found that the number of source documents from which citations were extracted did not influence performance noticeably, except when the number was either very low or very high. The number of seed documents used showed the same pattern. It is therefore concluded that a fairly low number of both quantities is sufficient for obtaining the best performance with citation searching as implemented in the boomerang effect. The number of seed documents needed is, however, much greater than a user could normally be expected to supply intellectually. Finally, it was investigated if an *expanded* citation index, containing the number of times a reference is mentioned in the full text, would be beneficial to retrieval performance. This could not be shown to be the case in the experiment. The main contribution of the dissertation is the exploration of these factors, and the boomerang effect as framework for experiments with both polyrepresentation and references and citations in best match IR. It is concluded that the limited effect of the factors may be due to the fact that the boomerang effect is at the un-structured pole of the polyrepresentation continuum, and recommended that future research investigates more structured approaches.

# Abstract in Danish

Formålet med afhandlingsarbejdet er at undersøge muligheden for at anvende referencer og citationer som en integreret del af automatiske indekserings- og genfindingssystemer, der opererer med videnskabelige fuldtekstdokumenter. Der er to hovedmotivationer bag afhandlingen: For det første viser de få studier fra både den brugerorienterede og den systemdrevne tradition, der har undersøgt referencer og citationer, generelt lovende resultater. For det andet, så begynder videnskabelige dokumenter i elektronisk form for første gang at blive tilgængelige i store mængder. Dette giver nye muligheder for at kombinere traditionelle automatiske indekserings- og genfindingsteknikker med udnyttelsen af referencer og citationer.

Både analytiske og empiriske undersøgelser udføres i afhandlingen for at undersøge, hvilke faktorer, der har indflydelse på, hvorledes automatiske indekserings- og genfindingsteknikker fungerer og klarer sig, når referencer og citationer er en integreret del af dokumentrepræsentationen af videnskabelige fuldtekstdokumenter. Det undersøges, *hvorfor* referencer og citationer kunne være anvendelige i genfindingsøjemed gennem en analyse af litteraturen om citationsmotiver og citationsadfærd. To synspunkter på forfatteres motiver for at citere andres arbejder identificeres: Et normativt og en socialkonstruktivistiskt. Det konkluderes, at den mest samlende teoretiske forklaring er Smalls (1978) opfattelse af referencer som 'concept symbols,' der står for en idé eller et begreb, der bliver brugt som del af et argument. I forhold til at anvende referencer og citationer i genfindingsøjemed er denne forklaring attraktiv, fordi, uanset om nogle referencer bliver udeladt, glemt eller er udvalgt ensidigt så fungerer de referencer der *faktisk udvælges* som symboler for et begreb. Dette kan forklare, hvorfor referencer og citationer kan være fordelagtige i genfindingsøjemed: Fordi referencerne repræsenterer begreber, når de bruges til dokumentrepræsentation, er de velegnede til genfindingsformål under forudsætning af, at brugerens informationsbehov kan udtrykkes i form af de samme begreber, dvs. som kernedokumenter. Analysen af tidligere forskning på området viste, at den mest almindelige brug af citationer netop var som kernedokumenter i en fremadrettet kædesøgning, hvor dokumenter, der refererer til kernedokumenterne identificeres som resultat. Den mest væsentlige faktor, der påvirker brugbarheden af referencer og citationer til genfinding, var muligheden for at udtrykke brugerens informationsbehov fyldestgørende i form af et eller flere kernedokumenter.

Metoden, der foreslås i afhandlingen – den såkaldte boomerangeffekt – er et forsøg på at fjerne kravet til brugeren om at hun skal angive kernedokumenter intellektuelt. I stedet identificeres disse automatisk ud fra brugerens forespørgsel i naturligt sprog. Inspireret af Ingwersens (1996) teori om polyrepræsentation udtrækker og prioriterer boomerangeffekten kernedokumenter fra en række kognitivt og funktionelt forskellige dokumentrepræsentationer, og den identificerer dokumenter, der citerer disse som resultat. Boomerangeffekten er placeret i den ustrukturerede ende af det polyrepræsentationskontinuum, der foreslås i afhandlingen som en udvidelse af teorien om polyrepræsentation.

Boomerangeffekten fungerer som ramme for de empiriske eksperimenter, der er gennemført som systemdrevne laboratorieeksperimenter med INEX testsamlingen (Gövert og Kazai, 2003). Hovedeksperimentet viste, at det er muligt at identificere relevante dokumenter via netværket af referencer og citationer med automatiske indekserings- og genfindingssystemer. Boomerangeffekten klarer sig lige så godt som en polyrepræsentationsbaseline uden citationer, men begge disse klarede sig dårligere end en traditionel baseline. En række faktorer blev undersøgt, for at analysere, hvorledes de påvirker boomerangeffekten. Det kan udledes, at antallet af dokumenter, der udtrækkes citationer fra, ikke påvirker resultaterne nævneværdigt, når blot der ikke er tale om et meget lille eller meget stort antal. Antallet af anvendte kernedokumenter viser samme mønster. Det konkluderes derfor, at et relativt lille antal af begge størrelser er tilstrækkeligt for at opnå de bedste resultater med citationssøgning som det ses implementeret i boomerangeffekten. Antallet af kernedokumenter er langt større, end man ville kunne forvente at en bruger vil kunne fremkomme med intellektuelt. Endelig blev det undersøgt, om et *udvidet* citationsindeks, indeholdende antallet af gange, hver reference er nævnt i den fulde tekst, kunne påvises at have en gavnlig effekt. Dette var ikke tilfældet. Afhandlingens hovedbidrag er undersøgelsen af disse faktorer og boomerangeffekten som instrument til eksperimenter med både polyrepræsentation og med referencer og citationer i vægtede genfindingssystemer. Det konkluderes, at de forholdsvis beskedne virkninger af faktorerne kan skyldes, at boomerangeffekten som den er implementeret i hovedeksperimentet, er placeret i den ustrukturerede ende af polyrepræsentationskontinuummet, og det anbefales at fremtidig forskning undersøger mere strukturerede tilgange.

Table of contents

# List of figures

# List of tables

# 1 Introduction

Increasingly, the activity of retrieving information is mediated by computerised systems designed for that purpose. In order for such an information retrieval system to be able to retrieve information desired by the user of the system, a representation of the user's information need must be matched somehow with representations of the documents contained in the system. Traditionally, retrieval of information from textual documents has been based on the matching of index keys from the documents with search keys of the same type from the user's request. The research field of information retrieval (IR) deals with facilitating this match and has developed a range of classic models for achieving this effectively and efficiently, e.g., the Boolean model, the vector space model and the probabilistic model (van Rijsbergen, 1979; Salton and McGill, 1983). Applications based on these models have been very successful and are now widely used in operational settings and in web search engines. It is, however, also generally recognised that there is room for improvements in terms of absolute performance, and the IR research community continues to develop and evaluate both new models and techniques, especially best match techniques, that attempt to rank the retrieved documents according to decreasing likelihood of relevance.

Traditionally, documents have been represented by index terms, either derived directly from the documents, or assigned by human indexers. Scientific documents can alternatively be represented by the *references* occurring in their bibliographies. This kind of representation was first proposed for scientific documents by Eugene Garfield (1955). The Science Citation Index® later developed by Garfield was made by inverting the references of scientific articles into an index, which facilitates the retrieval of documents that cite a given reference in their bibliographies (Garfield and Sher, 1963; Garfield, 1964). To this day the Science Citation Index® has remained operational as an alternative to the conventional indexing and abstracting services. Citation indexing can be regarded as an automatic indexing technique. However, relatively little interest has been shown in integrating citation indexing into IR research and into the fully automatic statistical indexing and retrieval techniques developed in parallel with the citation indexes (Spärck Jones, 2003).

The purpose of the dissertation is to investigate the use of references and citations as an integrated part of automatic indexing and retrieval techniques. The motivation behind this is twofold: *Firstly*, the few scattered studies that have investigated references and citations in IR have generally shown promising results. They indicated that the utilisation of references and citations in IR has the potential for improving retrieval performance. For instance, Salton (1971) found in a laboratory experiment that citation data were generally useful as indicators of document content, and that they were comparable in retrieval effectiveness to conventional term-based representations in a best match system. Another example is the work by Pao (1993), who in a field study found a much higher proportion of relevant documents among those documents that were identified by the combination of citation searches and conventional online searches.

*Secondly*, the full text of scientific documents in large quantities is for the first time becoming available in electronic form. This is a result of the change into an electronic production process, which has been completed in the last decade by the vast majority of the publishers of scientific documents. The change has been facilitated by the Standard Generalized Mark-up Language (SGML) and the development of standard Document Type Definitions (DTD) like ISO 12083:1994 for facilitating authoring, interchange and archiving of publications like books and journals. The introduction of a less complicated version of SGML called XML (eXtensible Mark-up Language) has lead to a much more widespread use of these kinds of formats. Documents formatted with such highly structured languages are particularly interesting because many specific types of information are marked up. This facilitates the automatic extraction of different types of information which can then be utilised for the representation of documents for IR purposes. A possibility with scientific documents, formatted in such a structured language, is to create citation indexes automatically from the mark-up. In the dissertation this is attempted as a basis for the investigation into the use of references and citations for IR purposes.

One problem with the use of citation indexes for information retrieval is that the user's information need usually has to be expressed in the form of a bibliographical reference. That is, not as a set of words that indicate what topic is sought for, but as a reference to a document which deals with the topic sought for and which subsequently may have been cited. While a user may be able to provide such 'seed documents' or 'seeds' for a

citation search, seed documents have not normally been available in the test collections used in most IR experiments. This is probably the main reason why citation indexing has not been exploited to a great extent in IR research. In addition, a user may not always be able to specify a seed document (for instance if she is not familiar with the topic), nor be able to specify one that is suitable for the purpose (Pao and Worthen, 1989). The dissertation proposes a method that, from the information need expressed in natural language, can exploit citation-based representations of the documents without the need for seed documents to be specified in advance. The so-called 'boomerang effect' automatically translates the natural language expression of the information need into references that are then used as weighted seed documents in a citation search[1].

It is likely that additional uncertainty will be introduced by such an automatic process, compared to a user's intellectual selection of seed documents. The intention behind the boomerang effect is to reduce this uncertainty by emphasising those citations that occur in the *overlaps* between the documents identified by a range of different cognitive and functional representations of the documents. This approach is based on the theory of polyrepresentation as put forward by Ingwersen (1992; 1994; 1996). The theory is part of the cognitive viewpoint in Information Science as represented, e.g., by the scholars B. C. Brookes, N. J. Belkin, M. De Mey, and P. Ingwersen.

The theory of polyrepresentation and the cognitive viewpoint also make up the theoretical background of the dissertation. The theory of polyrepresentation is holistic in the sense that it attempts to encompass both system-oriented and user-oriented approaches to IR research, and to create a robust framework for research through integration of both types of approaches into a consistent cognitive framework.

From a cognitive viewpoint the evaluation of IR techniques should ideally be carried out with real end users instead of in a laboratory setting. Unfortunately the boomerang effect as tested in the dissertation could not be developed to a state where users could interact with it because of lack of time. The experiments in the dissertation are therefore system-oriented because all tests were carried out in the laboratory setting without the

[1] The origin of the name is explained in Chapter 5 below. The boomerang effect has previously been presented in Larsen and Ingwersen (2001), Larsen (2002) and Larsen and Ingwersen (2002).

involvement of end users. However, even though the experiments and their evaluation are system-oriented all the design decisions, choice of document representations etc. are inspired by the theory of polyrepresentation. The test collection created by one of the most recent IR initiatives, the INEX test collection[2], was used for the main experiment. The document corpus in the INEX test collection is well-suited for the dissertation because it consists of 12,107 scientific documents in full text formatted in XML, including all their references.

## 1.1 Objectives of the dissertation

The main objective of the dissertation is to investigate the use of references and citations as an integrated part of automatic indexing and retrieval techniques operating on scientific full text documents. The dissertation attempts to contribute to the field of IR research by:

- Analysing *why* references and citations might be useful representations in IR,
- Analysing *how* references and citations have been utilised in IR previously, and
- Examining how a citation index might be constructed by automatic rather that intellectual methods from the full text of scientific documents.

The aim of the analyses is to identify factors that may affect the use of references and citations in IR and may need to be taken into account when utilising them as part of automatic indexing and retrieval techniques. Some of the factors that can be operationalised as variables can then be investigated empirically within the boomerang effect using the constructed citation index. The purpose of the experiments is to contribute knowledge about:

- Whether it is at all possible to retrieve relevant documents through the network of references and citations without specifying seed documents in advance,
- How the variables affect the operation of citation searches using a best match IR technique, and
- Which combination of variables that provides the best overall performance of citation searching.

---

[2] The Initiative for the Evaluation of XML retrieval (http://qmir.dcs.qmul.ac.uk).

In addition, because the approach to reducing uncertainty in the automatic selection of seed documents is based on the theory of polyrepresentation, a simple implementation of the theory, which does not utilise the citation index, is also tested in the experiments as a baseline. The purpose of the inclusion of this polyrepresentation baseline is to contribute knowledge about:

- How the polyrepresentation baseline behaves and performs in comparison to the boomerang effect.

Finally, a baseline that represents a standard best match IR technique is also included in the experiments, which may contribute knowledge about:

- How the boomerang effect and polyrepresentation baseline behave and perform in comparison to standard best match IR techniques.

## 1.2 Research questions

The overall research question investigated in the dissertation is:

> *Which factors affect the behaviour and performance of automatic indexing and retrieval techniques given that references and citations are an integrated part of the document representation of scientific full text documents in the IR system?*

The identification of possible factors is addressed analytically in a review of the literature concerning citation theory and the literature where references and citations have been utilised for IR purposes. Some of the factors were operationalised in a best match version of the boomerang effect, which served as a framework for the experiments. A number of specific research questions were investigated empirically, mainly by studying the factors' effect on performance:

1. *Does the best match boomerang effect achieve a similar level of performance compared to what is obtainable with a polyrepresentation baseline and a best match baseline?*

The question is a general one and seeks to investigate if the boomerang effect achieves a similar level of performance as the two baselines. Through investigating this question it is also hoped that much can be learned about the functioning of the boomerang effect

and the behaviour of references and citations in relation to automatic indexing and retrieval techniques.

2. *Are there significant differences in IR characteristics and performance between individual representations of the scientific full text documents?*

Because the boomerang effect needs different cognitive and functional representations in order to function such representations were generated of the documents. Differences among each of these individual representations are interesting, because such differences may need to be taken into account in future studies which incorporate implementations of the theory of polyrepresentation. In the present dissertation the individual representations form the source of the boomerang effect and the baselines. Differences among them will influence the behaviour of both the boomerang effect and the baselines, and knowledge of such differences may illuminate the obtained results. The differences in IR characteristics examined in the dissertation include for example the number of documents in which each representation can be identified and extracted, the number of documents retrieved on average per representation, as well as the performance of each representation separately.

3. *Does an increase of the number of source documents in the simple polyrepresentation baseline improve performance?*
4. *Does performance improve by increasing the number of documents from which references are extracted in the best match boomerang effect?*

One of the factors tested in the boomerang effect is the number of documents from which references are extracted as sources for the automatic weighting and selection of seed documents. Research question 4 investigates if the performance of the citation search strategy implemented in the best match boomerang effect can be improved by increasing the number of source documents. Similarly, research question 3 investigates if increasing the number of source documents can improve the performance of the polyrepresentation baseline.

5. *Can better performance be obtained by reducing the number of seed documents in the best match boomerang effect to the seed documents with the highest weights?*

The question investigates if better performance can be obtained by using a limited set of seed documents rather than all those identified automatically by the boomerang effect. The seed documents with the lowest weights might for instance act as noise.

6. *Can better performance be obtained by using an expanded citation index as basis for the weighting and selection of seed documents in the boomerang effect compared to a flat citation index?*
7. *Can better performance be obtained by running the citation queries against an expanded citation index in the boomerang effect compared to a flat citation index?*

One of the novel possibilities offered by scientific full text documents in electronic form is that an expanded citation index can be constructed. Such an expanded citation index would register, not only the references in the documents, but also the frequency with which they are mentioned in the full text. Such an expanded citation index might result in better performance of citation searching compared to a conventional (flat) citation index as proposed by Herlach (1978). Both a flat and an expanded citation index were constructed for use in the boomerang effect. Research question 6 investigates whether the use of the expanded citation index for the automatic weighting and selection of seed documents in the boomerang effect can improve performance. Research question 7 investigates if increased performance can be obtained by running the citation queries against an expanded citation index.

## 1.3 Structure of the dissertation

The dissertation falls into two main parts: a theoretical and an empirical one.

*The theoretical part* consists of Chapters (1), 2, 3, 4 and 5. Together they provide the background information and literature reviews in support of the boomerang effect. Chapter 2 presents the three main lines in IR research, the system-oriented tradition, the user-oriented tradition and the cognitive approach. Chapter 3 presents the theory of polyrepresentation, and discusses its relevance to the dissertation. In Chapter 4 it is analysed why references and citations might be useful as representations in IR, and earlier research on the use of references and citations in IR is reviewed. Chapter 5 presents two versions of the boomerang effect: a Boolean version and a best match version, as well as a pre-experiment with the Boolean version in which the basic functioning of the boomerang effect was examined.

*The empirical part* consists of Chapters 6, 7 and 8, which give an account of the empirical experiments. One should note the fact that the boomerang effect was developed and implemented in three rounds: 1) In the pre-experiment which tested the basic functioning of the boomerang effect in a Boolean setting. The details of the experiment and a discussion of the experiences gained are given in Chapter 5. 2) In a preliminary submission to the INEX initiative from the TAPIR group[3] (referred to as the official INEX submission), because active participation in INEX was a prerequisite for access to the test collection. 3) In the main experiment, where considerable improvements were achieved over the official INEX submission by studying (or tuning) the variables of the boomerang effect. The details of the preliminary INEX submission, as well as of the best match boomerang effect in the main experiment are given together in Chapters 6-8 as they are very similar in execution.

Chapter 6 gives details of the test data and methods used in the experiments, including the INEX test collection and implementation of the best match boomerang effect and the baselines. Chapter 6 includes a description of the InQuery IR system, which was used as basis for the construction of the best match boomerang effect[4]. Chapter 7 presents the empirical results of the research questions. The discussion of the results is placed separately in chapter 8 because several of the research questions interact with each other. The analytical results are also discussed in Chapter 8.

---

[3] The dissertation is part of the TAPIR (Text Access Potentials for interactive Information Retrieval) project headed by Professor Peter Ingwersen (See http://project.dbit.dk/tapir/).

[4] The InQuery IR system has been developed at the Center for Intelligent Information Retrieval (CIIR), Computer Science Department at University of Massachusetts, and has kindly been made available to the TAPIR group. InQuery is described in detail in Chapter 6 below.

Chapter 9 summarises the main conclusions and outlines the contributions made by the dissertation.

Chapter 10 contains the bibliographic references, and is followed by a list of the abbreviations used in the dissertation and the appendices.

# 2 IR research approaches

In this chapter the three main approaches to IR research are presented, and the dissertation is placed in the context of these approaches.

The actual number and character of research traditions one may identify in IR research depends very much on the point of view from which the field is analysed. A great deal of the research in IR has been carried out within what may be characterised as the system-driven tradition. Within this tradition there is a tendency that writers mostly focus on this tradition alone. Other approaches are dealt with in passing or ignored entirely. For example, in their textbook *Modern Information Retrieval* Baeza-Yates and Ribeiro-Neto distinguish between two different views of IR, a computer-centred one and a user-centred one, but only the former is dealt with in the book (1999). Other writers analyse more than one approach, e.g., Ellis (1992; 1996; 1998) argues that there are two main research approaches in the field, the archetypal approach and the cognitive approach. The former corresponds to the system-driven tradition, and the latter consists of a wide range of IR research that focuses on more complex and dynamic interactions between user and system.

The present dissertation is inspired by a broad conception of the field, which can accommodate all of these different views:

> "Information retrieval is concerned with the processes involved in the representation, storage, searching and finding of information which is relevant to a requirement for information desired by a human user." (Ingwersen, 1992, p. 49)

To this conception of IR one might add processes like 'filtering', 'visualisation' and, at the end 'through interaction'. This extension of Ingwersen's understanding is due to the developments of IR-related IT in the last decade.

In line with this extended view, the dissertation takes a correspondingly broad view on the research traditions, viewing IR as consisting of three lines of research: The system-driven tradition, the user-oriented tradition and the cognitive approach. The system-driven tradition is the oldest and largest in terms of research volume. The user-oriented tradition is younger, but growing. The cognitive approach represents a relatively novel

and quite holistic perspective on IR research. Figure 2.1 shows the main actors and components in IR, and will be used for the characterisation of the three main IR approaches.

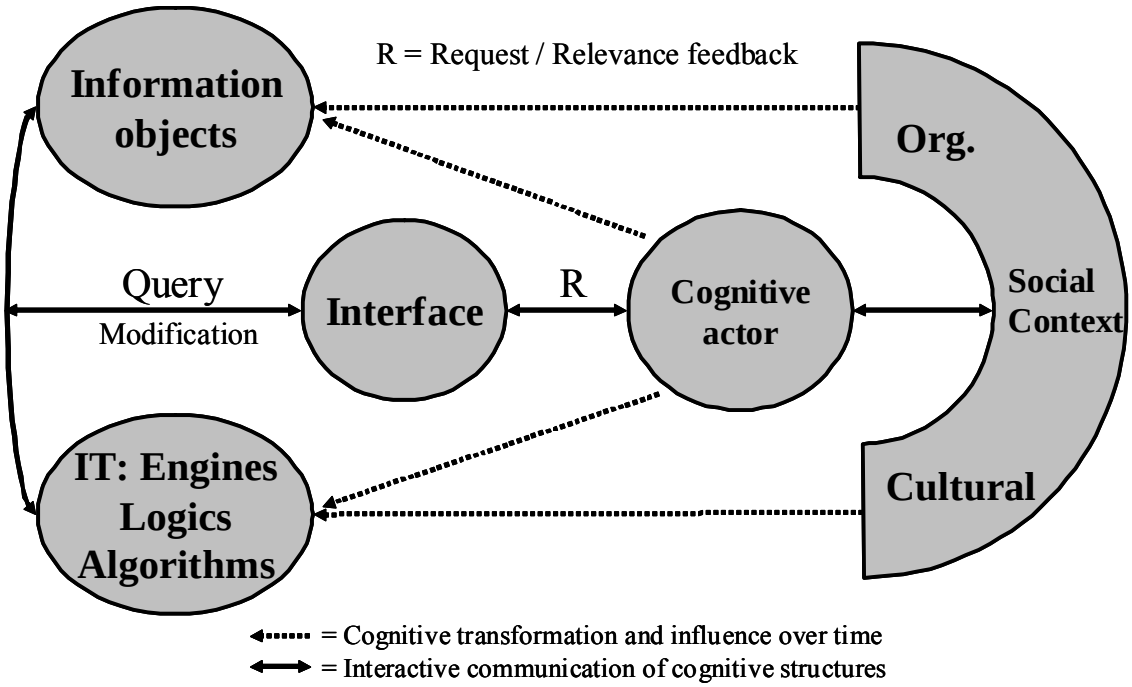



**Figure 2.1. Simplified model of IR interaction. Revised from Ingwersen (1992) and Ingwersen (1996). (Courtesy of Peter Ingwersen, 2003. To appear in a modified form in Ingwersen and Järvelin ([2004])).**

## 2.1 The system-driven tradition

The main purpose of the system-driven tradition is to design and test better IR algorithms and systems. An important characteristic of the system-driven tradition is that the algorithms are not only designed, but also tested for their effectiveness. Indeed, the question of how to evaluate retrieval is one of the determining characteristics which separate the three research approaches.

The system-driven tradition is the oldest of the three and originates in the comparative tests of indexing languages in the late 1950s and early 1960s (Ellis, 1996). The two main tests were the Cranfield I and Cranfield II experiments (Cleverdon, 1960; 1962; Cleverdon, Mills and Keen, 1966; Cleverdon and Keen, 1966). The Cranfield

experiments were very influential because they "…established the principle that arguments about the relative merits of different retrieval system designs had to be empirically grounded." (Ellis, 1996, p. 19). This represents a change from earlier where IR system design was mainly a philosophical and speculative endeavour. The approach developed as a framework for obtaining empirically grounded arguments stems directly from the Cranfield II experiment, and is consequently often referred to as the Cranfield model. The model is based on the principle of test collections. A test collection consists of a document corpus, a set of (usually quite well-defined and topical) requests[5], and a set of relevance assessments that identify documents from the corpus that are relevant to each request. The main focus of the research in the system-oriented tradition has been to develop theories and methods for generating representations of the documents and the requests, and effective algorithms for the matching of these representations within the Cranfield model. The main forum for the system-driven tradition today is the annual Text REtrieval Conferences (TREC), which was begun in 1992 (See, e.g., Harman, 1993).

Thus, the system-driven tradition is concerned with the components on the outermost left-hand side of the model in Figure 2.1 (The information objects, queries, and IT: Engines, logics, algorithms), and the relations between them. Because no potential users are involved in the experiments (other than perhaps as relevance assessors) the model is also referred to as the Laboratory Model, emphasising that the experiments carried out within the model take place in the laboratory rather than in an operational environment. Figure 2.2 provides a detailed overview of the Laboratory Model and the components it is concerned with. The middle of the figure (beginning with the documents and requests and ending with the [retrieval] result) shows the components of the IR system and corresponds to the left-hand side of Figure 2.1. If an interface was added to the components in the middle of figure, this would what is needed for a functioning IR system that would allow users to interact with the system. The components included in the evaluation are shown in the lightly shaded 'horse-shoe' area at the top, left side and bottom of the figure. Users have only recently been involved within the Laboratory Model (See, e.g., Over, 1997). This is indicated in the darkly shaded area to the right of the figure.

---

[5] These are also often called 'topics'.

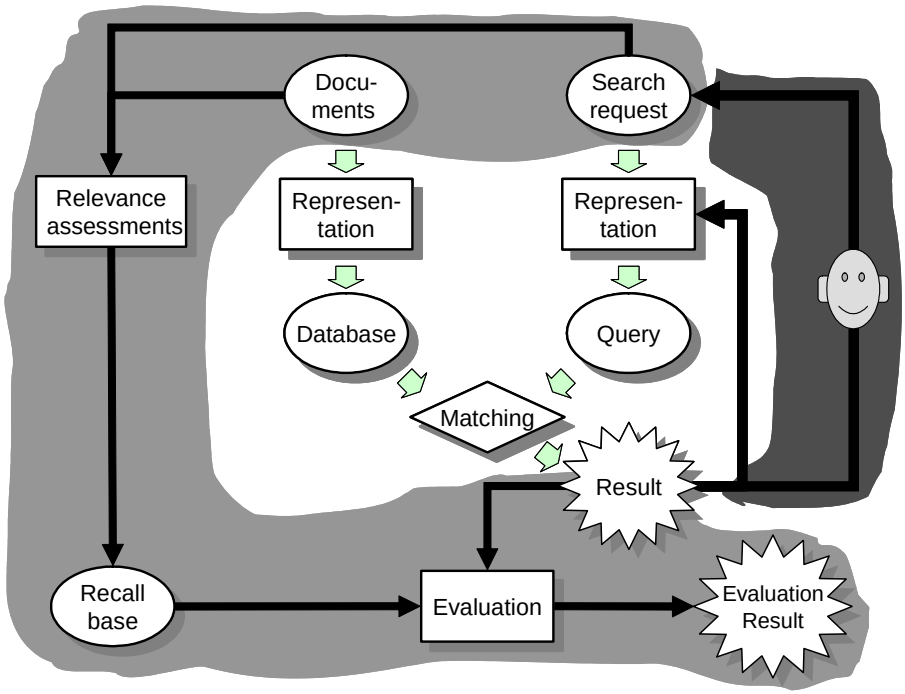



**Figure 2.2. The Laboratory model schematised. After Kekäläinen and Järvelin (2002a, p. 255)**

Figure 2.2 shows how representations are generated of both documents and requests, resulting in a searchable database as well as queries, which are expressions suitable for matching against the database. By running each query against the database the matching algorithm then produces a retrieval result. The matching algorithm typically produces a ranked output where the retrieved documents are ranked according to decreasing likelihood of relevance (Belkin and Croft, 1987). An important point made obvious by Figure 2.2 is that that the process of generating representations of documents and queries must produce representations that are of the same type – that is, any features that one might wish to incorporate or exploit from either documents or requests must be supported by the other. For instance, if a system designer wishes to allow phrase searching in the query language, the database index must contain some sort of postings information to support this. The main goal of the dissertation is to investigate the potentials of references and citations as an integrated part of the document representation of scientific full text documents in automatic indexing and retrieval techniques. Therefore methods must be identified to represent the requests in a way that is *matchable* with citation-based representations of the documents – for instance in the

form of seed documents. The boomerang effect presented in Chapter 5 is an attempt to achieve this. In terms of Figure 2.2 the dissertation circumscribes the standard Cranfield model, i.e., the lightly shaded horseshoe-formed area.

In the system-driven tradition the quality of the retrieval result produced by a particular IR technique is evaluated by its ability to retrieve relevant documents from the document corpus. As shown in Figure 2.2 the relevance assessments from the test collection are used to form a recall base for each request. In the evaluation the recall base is compared to the retrieval result to compute standard performance measures like recall and precision which are averaged over requests (Cleverdon and Keen, 1966, p. 31. See also Section 6.4 below). The requests have typically been topical in nature and the relevance assessments carried out by human assessors with expertise in the topic of the request. Therefore the type of relevance reflected by the relevance assessments used in the system-oriented tradition can be characterised as *topical relevance* or simply *topicality* (Saracevic, 1996). Among other issues, the use of this kind of relevance assessments as the basis of evaluation in the Laboratory Model has attracted strong criticism. Kekäläinen and Järvelin investigated the objections against the model one by one and concluded that although it can be improved in certain aspects "…the Laboratory Model is, while clearly limited in scope, not as notorious as criticized when used as a model for developing and evaluating IR algorithms for interactive IR systems." (2002a, p. 267).

The main justification for the Laboratory Model is that the representations of documents and queries, as well as the relevance criteria used, are of similar nature because they are topical. The main part of the criticism levelled at the model concerns questions that are outside the scope of the model: to make improved algorithms. Kekäläinen and Järvelin pointed out that while this may be a restricted and narrow view; other broader evaluation scenarios are needed to design IR systems that take account of, e.g., the information seeking strategies of users over several sessions and their use of IR systems in their real life work tasks. Research into those types of questions has been carried out in the user-oriented tradition and in the cognitive approach discussed below.

The major achievements of the system-driven tradition include three statistically-based IR models (the Vector Space Model (Salton and McGill, 1983) and the Probabilistic Model (Robertson and Sparck Jones, 1976; van Rijsbergen, 1979)), as well as the use of

term frequency information in these models. The InQuery retrieval engine applied in the experiments discussed in this dissertation belongs to the latter type of IR model. It uses a modification of the probabilistic Bayesian belief network model to create an Inference Network Model (Turtle and Croft, 1990). The successful best match IR systems, which can produce a ranked output according to decreasing likelihood of relevance, are based on these ideas. These systems use representations generated from the possible indexing keys, like the words in the documents, which are then weighted according to their frequency in the document and across the collection according to a weighting scheme (Spärck Jones, 2003). The so-called *tf*idf* weighting formula assigns a weight to each index key in the document according to its frequency in the document (*tf*), and the inverse frequency of the index key in the collection (*idf*). (See, e.g., Salton and Buckley, 1988, or Section 6.2.1 below).

## 2.2 The user-oriented tradition

The research carried out within the user-oriented tradition focuses on the behavioural and psychological aspects that affect the user's interaction with IR systems. Compared to the system-driven tradition, IR is seen in a broader perspective as a problem solving and goal oriented, interactive process. The general aim in the tradition is to improve IR effectiveness by studying individual users' searching behaviour in empirical, real-life investigations. The results produced by the tradition include complex models of information behaviour made by studying common patterns between searchers of information (Ingwersen, 1992).

The user-oriented tradition has two sub-traditions: the operational R&D approach, and the information seeking research (Ingwersen, 1996, p. 12-13). The *operational R&D approach* has mainly been concerned with individual users and their interaction with online systems and web search engines through some form of interface. In Figure 2.1 the attention of the operational R&D approach has been on the cognitive actors (be it intermediaries such as librarians or end users) and their interaction with one another and/or the interface. Typical examples of such studies include, e.g., the large-scale empirical investigations by the Saracevic and Kantor group (Saracevic et al., 1988; Saracevic and Kantor, 1988a; 1988b) and later by Saracevic et al. (Saracevic, Mokros and Su, 1990; Saracevic et al., 1991). In the mid-1990s Spink (1996; 1997) also investigated end users as well as feedback utilisation during standard online searching,

including longitudinal session periods. Single session investigations of end user online searching was followed up by Su (1994) on the perception of recall versus precision by searchers, and lately Ellis et al. (2002) as well as Wu and Liu (2003) re-looked into intermediary activities, following up on the early Ingwersen study from 1982. Longitudinal investigations seem mainly to belong to the information seeking research or the cognitive approach.

The *information seeking research* takes a broader perspective and carries out investigations of information behaviour in organisational, social and cultural contexts to arrive at general models of interaction between actors and over time. Examples include, for instance, Kuhltau's information search process model that, based on a number of longitudinal studies, shows that students' and library users' information search behaviour consist of a number of different stages (Kuhlthau, 1991). The force of her investigations and model is the inclusion of emotional characteristics of searchers and the conception of levels of uncertainty and doubt associated with the searching stages. In Figure 2.1 the attention of the information seeking research is on the right-hand side, at the cognitive actors and their context.

A large variety of themes have been touched upon by the common user-oriented tradition. These include, e.g., analyses of the nature of the information need (Taylor, 1968), the different types of information needs that can be identified (Ingwersen, 1996), as well as a large body of research into user-oriented relevance since 1990. (See, e.g., Schamber, Eisenberg and Nilan, 1990; Saracevic, 1996; Borlund, 2000a; Cosijn and Ingwersen, 2000). Evaluation of IR systems in the user-oriented tradition takes a correspondingly broader view: "The main purpose of this type of evaluation is concerned with how well the user, the retrieval mechanism, and the database interact extracting information, under real-life operational conditions. In this approach the relevance judgements have to be given by the original user in relation to his or her personal information need which may change over session time." (Borlund, 2000a, p. 57).

Thus, as a whole the user-oriented tradition tackles IR research from the perspective of user(s), hence the name. The actual systems involved and the characteristics of their algorithms, representation techniques, etc. have rarely been a concern of the tradition:

> "…the objectives for research, and the models and results published, omit all or several system components….The user-oriented approach does not, within the individual projects, concern itself with the problems of different text representation and IR technique issues. In the traditional IR approach [i.e., the system-driven tradition] the users and intermediaries hardly exist. Similarly, the user-oriented approach in general takes system components to be constants, rarely linked to the human ones. This is presumably a natural consequence of the real-life R&D environment which either involves printed retrieval tools with poor and similar access possibilities or exact match online retrieval only….As a consequence, the user-orientation demonstrates a similar restricted view of the total IR situation, and has difficulty in contributing to more overall IR theories." (Ingwersen, 1992, p. 84)

I believe that these observations made a decade ago are still valid although the research has moved into the Web environment. The operational engines are commonly taken for granted.

The IR research that attempts to bridge the two traditions and integrate the goals, methods and results from both can be considered to belong to the cognitive approach reviewed below (Ingwersen, 1996).

## 2.3 The cognitive approach

It is apparent from the two previous sections that the system-driven tradition and the user-oriented tradition provide two very different perspectives on IR. The former is concerned with the development of representation techniques and matching algorithms to be implemented in IR systems, but without much concern for the real life situations in which users seek information. The latter mainly deals with users' behaviour when seeking information, but it is not much concerned with the features of the IR systems from which the information is retrieved.

The *cognitive approach to IR research* represents an attempt to bridge the gap between the two traditions, and to create a framework where the theories, methods and results from both can be integrated into a coherent framework. That is, to view all elements in Figure 2.1 including the left, middle and right-hand side from a common perspective in

IR research. Such a perspective that can integrate the system-driven tradition and the user-oriented tradition is provided by the cognitive viewpoint in Information Science, which states that "…any processing of information, whether perceptual or symbolic, is mediated by a system of categories or concepts, which for the information-processing device, are a model of the world." (De Mey, 1980, p. 48). Ingwersen (Ingwersen, 1992, p. 146-156) gives examples of how each of the components in Figure 2.1 can be viewed as the result of transformations of world models or knowledge structures of the actors involved. All the basic components in IR can thus be discussed and analysed from the same theoretical perspective creating new possibilities for more coherent, but also more complex research.

Not all research that integrates the two main traditions adheres explicitly to the cognitive approach. The prototype IR systems from the mid-80s, which incorporated artificial intelligence techniques in an attempt to build expert systems that could act as intermediaries between the IR system and the user are examples of research that is not explicitly carried out within the cognitive approach. However, systems like the $I^3R$ system (Croft and Thompson, 1987) and the Bookhouse system (Mark Pejtersen, 1989) integrate knowledge about the information seeking behaviour with different types of document representations and matching algorithms. Examples of research that have applied the cognitive approach directly is, e.g., Wilson's work on a cognitive approach to information seeking behaviour and information use (1984), Borgman's investigation of individual differences between users of IR systems (1989), Bruce's proposal of a methodology for controlling situational dynamism in users' relevance assessments in interactive IR (1994), and Borlund's dissertation work on establishing an alternative approach to the evaluation of interactive IR systems (2000a).

Evaluation of IR systems in the cognitive approach is difficult because of the holistic nature of the approach: on one side it is desirable to maintain the experimental control and relative efficiency of the system-driven tradition, on the other this is clearly too limited in light of the experiences from the user-oriented tradition. Borlund remarks that the latter's approach to evaluation as described in Section 2.2 seems ideal for the evaluation of interactive IR (IIR) systems, except for the lack of control and cost associated with involving real users with real and variable information needs in the experiments (Borlund, 2000a, p. 58). The analytically and empirically grounded 'IIR evaluation package' proposed by Borlund can be regarded as a constructive attempt to

combine the desirable properties from both traditions within the cognitive approach. The IIR evaluation package consists of three parts:

1. "the proposal of a set of components which aims at ensuring a functional, valid, and realistic setting for the evaluation of IIR systems;
2. empirically based recommendations for the application of the proposed sub-component, the concept of a simulated work task situation; *and*
3. the proposal of alternative…performance measures which are capable of bridging the interpretative distance between objective and subjective types of relevance involved in the evaluation of IIR systems, as well as managing non-binary relevance assessments." (Borlund, 2000a, p. 77).

Together the three parts ensure that the *realism* of the user-oriented tradition and the *control* of the system-driven tradition can be combined in a consistent and reliable way. Thereby Borlund roots the IIR evaluation package firmly in the cognitive approach, and thereby demonstrates the viability of this approach to IR research.

An example of theory building within the cognitive approach is the theory of polyrepresentation by Ingwersen (Ingwersen, 1996). In short, the theory of polyrepresentation hypothesises that overlaps between different cognitive representations of both users' information needs as well as documents can be exploited for reducing the uncertainties inherent in IR, and thereby improve the performance of IR systems. Two, or more, different cognitive representations pointing at the same documents are regarded as multi-evidence of those documents being relevant. This suggests applying a principle of 'intentional redundancy' (Ingwersen, 1994) with the purpose of reducing the uncertainties by placing emphasis on *overlaps* between representations. Better results are expected when cognitively unlike representations are used, e.g., the document title (made by the author) vs. intellectually assigned descriptors from indexers. The dissertation is based on the cognitive approach and the theory of polyrepresentation, and more details about them are given below in Chapter 3.

## 2.4 The dissertation in relation to the traditions

The ideas investigated in the dissertation are inspired by the cognitive viewpoint in general and the theory of polyrepresentation in particular. Therefore the intention in the

original dissertation proposal was to carry out a fully-fledged experiment within the cognitive approach. That is, testing the potentials of references and citations as an integrated part of a best match IR system in an interactive setting involving end users. However, quite early in the process it became clear that this was too ambitious within the given time frame, especially the construction of an interactive IR system prototype that would be appropriate for such experiments. Scholars experienced with the execution of empirical IR experiments offered the kind advice that the setting up of a suitable test collection and a basic implementation to support citation searching would be quite a challenge in itself.

The empirical experiments carried out in the dissertation are therefore system-oriented because all tests were carried out in a laboratory setting without the involvement of end users. The investigated research questions are firmly rooted in the Laboratory Model, because they are oriented towards the development of better algorithms by the study of an alternative type of representation of both documents and requests. However, even though the experiments and their evaluation are system-oriented all the design decisions, choice of document representations etc. are inspired by the theory of polyrepresentation and the cognitive viewpoint. The dissertation work can therefore be seen as the first step towards setting up an environment for the TAPIR project within which we may study users as they interact with complex IR systems thus integrating both sides of Figure 2.1.

# 3 The theory of polyrepresentation

The present dissertation is based on the cognitive viewpoint and the theory of polyrepresentation in particular. The aim of this chapter is to present the cognitive viewpoint in brief and to present and discuss the theory of polyrepresentation and how they relate to the dissertation.

## 3.1 The cognitive viewpoint

The origin of the cognitive viewpoint cannot be traced back to a single event or person who has founded it. Rather, a number of publications emerged from the mid-70s and onwards that called for or proposed that a cognitive view of Information Science might be beneficial (Belkin, 1990). Part of the motivation behind these proposals was a wish to create an alternative scope for Information Science research, in particular an alternative to the system-driven tradition (Belkin, 1984). Although there was not an exact agreement on the precise "…definition of what such a view is, or what it entails, there was a consensus of the meaning common to them all." (Borlund, 2000a, p. 12). Borlund (2000a, p. 11-19) reviews the cognitive viewpoint by analysing the work of four scholars that can be seen as personifying it: B. C. Brookes, N. J. Belkin, M. De Mey, and P. Ingwersen. Her main conclusions are rendered below.

*B. C. Brookes* is one of the earliest advocates of the cognitive viewpoint, and his main contribution is his 'fundamental equation of Information Science' (Brookes, 1975a). The equation (1975b; 1977; 1980) was intended to function as a catalyst for posing questions in order to understand and uncover the scope of the field of Information Science, rather than as an exact mathematical formula. The equation illustrates how knowledge structures are affected and modified by external information. As Belkin (1990) points out, this demonstrates the power of the cognitive viewpoint because it emphasises the involved knowledge structures and the interaction between them. In Brookes' view "…the interpretation of the fundamental equation is the basic research task of information science…" (Brookes, 1975a, p. 117)

Inspired by Brookes, *N. J. Belkin* continues the search for an understanding of the fundamental problem of Information Science, by regarding it as "…the effective transfer of desired information between human generator to human user…" (Belkin and Robertson, 1976, p. 197; Belkin, 1977, p. 187). The focus on the information processing and transfer between human actors is expressed clearly in his famous ASK hypothesis, which he defines to be the recognition of an anomaly – or information gap – by the recipient in his/her state of knowledge (Belkin, 1980; 1982). This 'anomalous state of knowledge' (ASK) results in a need for information in order to reduce uncertainty or to solve a particular problem. Borlund notes that it is due to the ASK hypothesis that the cognitive viewpoint makes a breakthrough in IR research, because this changes the idea of an information need from a static concept into a "…user-individual and potentially dynamic concept…" that has proven particularly useful for the user-oriented tradition (Borlund, 2000a, p. 14).

In parallel with Brookes and Belkin, *M. De Mey* connects what he regards as a novel and powerful movement in many research fields, the cognitive paradigm, to the field of Information Science (De Mey, 1977; 1980). His contribution to the cognitive viewpoint in Information Science is mainly epistemological, e.g., by providing a four-stage model of how ideas about information processing have advanced[6]. The stages range from the *monadic* (where information is handled in isolation) through the *structural* and the *contextual*, to the *cognitive/epistemic* stage (where information is handled in terms of individual world models) (De Mey, 1980, p. 80). Borlund (2000a) remarks that the fourth and final stage corresponds to how the cognitive viewpoint can be regarded as an alternative to the system-driven tradition because most human information processing, including information need formation, take place at this level (Ingwersen, 1992, p. 23). In contrast, the system-driven tradition can be said to have been concerned mainly with the first stage.

*P. Ingwersen's* contribution to the cognitive viewpoint can, according to Borlund (2000a, p. 16-17), be divided into two types: The *first* type of contribution consists of further developments of the works of fellow scholars. This includes, e.g., the suggestion of improvements to Brookes' equation (Ingwersen, 1984; 1992), and the addition of

[6] De Mey (1980) ascribes the origin of the model to Michie (1974).

specifications or conditions to the cognitive viewpoint, for instance by pointing out that there is a perceived work task (or interest) behind an ASK, and by extending Belkin's 1980 model of the cognitive communication system. Figure 2.1 above is a further development of Belkin's model. Fundamentally Ingwersen sees information seeking and interactive IR as forming a part of processes of cognition. The *second* type of contribution consists of demonstrations of the applicability of the cognitive viewpoint, e.g., in an empirical investigation of the transfer processes involved in reference work in public libraries, or in bibliometric and scientometric work on exploiting representations of different cognitive origin in online publication and citation analysis (Hjortgaard Christensen and Ingwersen, 1996). The development of the theory of polyrepresentation (Ingwersen, 1994; 1996) discussed below can also be regarded as a direct consequence of the applicability of the cognitive viewpoint to theory building. In Borlund's view, which this dissertation supports, the further developments of the cognitive viewpoint by Ingwersen "…illustrate his holistic view of the IR interaction scenario within the field of Information Science. With his holistic cognitive view Ingwersen emphasises how each of the involved cognitive actors (e.g., the information generator, the information representer, the intermediary, and the information recipient/user) are of equally importance in order to achieve successful and optimal IR." (Borlund, 2000a, p. 17).

Overall, the cognitive viewpoint provides a much more comprehensive view on IR than the system-driven and user-oriented traditions described in Chapter 2 above. The main advantage of the cognitive viewpoint is that it attempts to integrate all actors in IR within the same theoretical framework, or as put by De Mey:

> "…any processing of information, whether perceptual or symbolic, is mediated by a system of categories or concepts, which for the information-processing device, are a model of the world." (De Mey, 1980, p. 48)

This understanding implies that the cognitive agents behind computer configurations (where the latter is seen as representing the formers' world views) are central to the information processing outcome, that is, the 'symbolic' processing. For IR research this has the consequence that more wide-ranging research questions can be addressed than if either the system-driven tradition or the user-oriented tradition is applied in isolation. Therefore, issues that may seem insignificant or trivial from either of these perspectives may gain new importance when combined in a coherent framework. Inspired by this, the dissertation attempts to work consciously with document representations that have different cognitive origins, and to combine these in a comprehensive manner. See

Ingwersen (2001) for a recent and more in-depth discussion of the cognitive IR theory, including criticism against its views.

## 3.2 The theory of polyrepresentation

In short, the theory of polyrepresentation "…hypothesises that overlaps between different cognitive representations of both users' information needs as well as documents can be exploited for reducing the uncertainties inherent in IR, and thereby improve the performance of IR systems." (Larsen and Ingwersen, 2002, p. 397). Ingwersen developed the theory of polyrepresentation through the 1990s. The theory is fully expanded in the *Journal of Documentation* article from 1996 (Ingwersen, 1996), which remains the main publication on the theory. 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 SIGIR (Ingwersen, 1994). A recent update to the theory appeared at the CoLIS4 conference (Ingwersen, 2002).

Ingwersen views all communication processes in IR as consisting of interchanges that take place at the *sign level*. When humans are part of the 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 in line with Brookes' conception outlined above in Section 3.1. When a machine is the recipient it may react on the received information, but only at the sign level and only by the responses it has been pre-programmed with. However, Ingwersen regards this pre-programming as human cognitive structures that have been embedded in the machine prior to the information processing. 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 generator to achieve communication at the cognitive level (Ingwersen, 1996, p. 6). This leads to two fundamental characteristics which are of importance to IR:

- "the uncertainties and unpredictabilities inherent in IR interaction;
- any presuppositions, meaning and intentionality underlying the communicated messages are vital but constantly lost" (Ingwersen, 1996, p. 8)

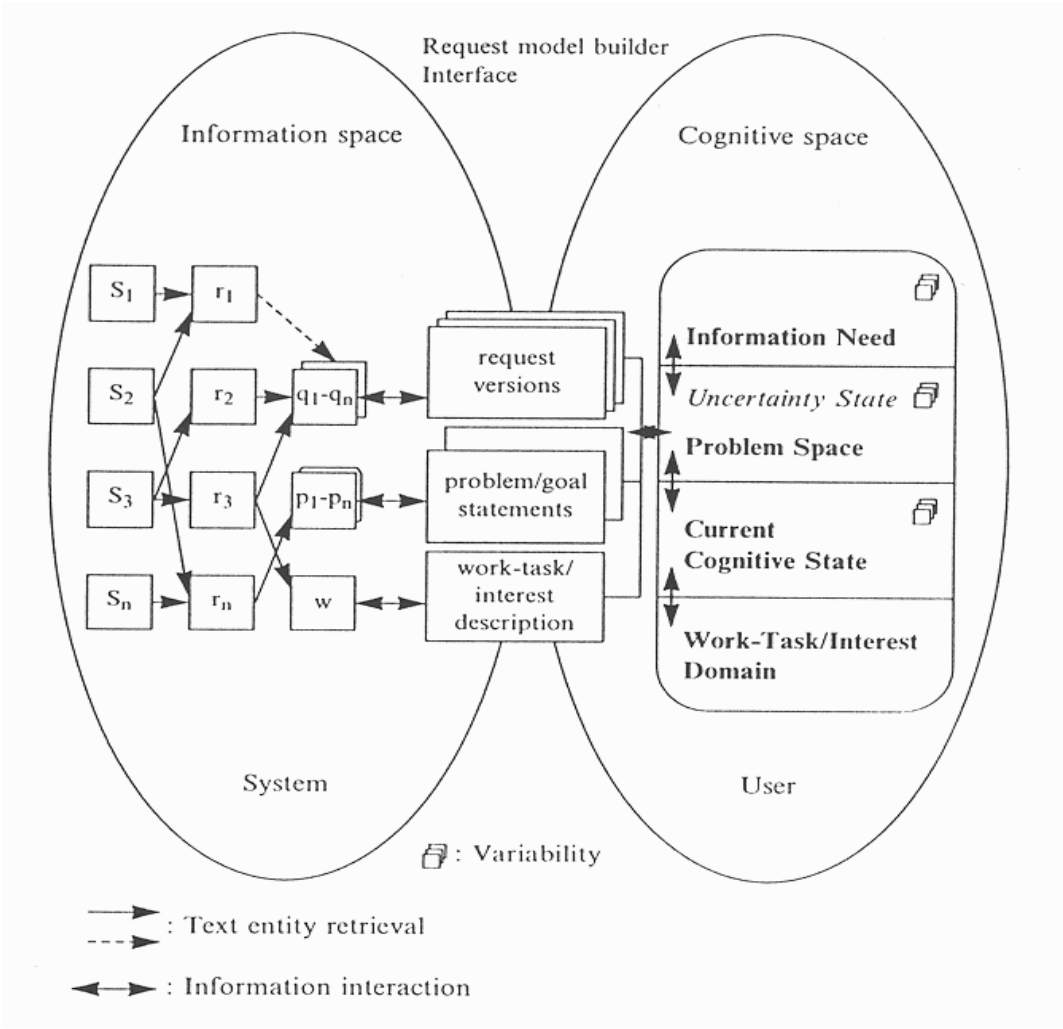



**Figure 3.1. The global model of polyrepresentation. (From Ingwersen, 1996, Figure 8, p. 37)**

Much research and development work have been done to reduce such uncertainties and unpredictability, e.g., by building controlled vocabularies like thesauri or by setting up extensive cataloguing rules. The theory of polyrepresentation represents an attempt to view the uncertainties and unpredictability as favourable to IR, and to exploit these actively. This is achieved by focussing on the cognitive structures and the representations that may be generated from these in both the cognitive space of the user *and* in the information space of the IR system. The inescapable and inherent uncertainties are part of all these representations. The main hypothesis in the theory of polyrepresentation is that the more cognitively and functionally different representations

that point at a particular document, the greater the probability that this document is relevant. Figure 3.1 illustrates the representations that may be generated from the cognitive space of the user and the additional possibilities of matching them with representations of the documents. In Figure 3.1 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. Each of these may give rise to various representations ($r_1 \ldots r_n$) that may be matched with different functional representations of the user's information need, either the work task description (w), problems statements ($p_1$-$p_n$) or requests ($q_1$-$q_n$).

In the theory of polyrepresentation all tangible representations involved in IR, (See Figure 2.1) are referred to as 'cognitive representations', because they in some way or other always arise from the activity of a human actor, and therefore are regarded as the transformed knowledge structures. The same representations are also referred to as 'functional representations' when several representations are derived from the *same* cognitive actor, and display characteristics that make it possible to distinguish between them. Examples of this are the title versus the abstract of a scientific article, or a description of the underlying work task versus a description of the present information need derived from a user. In my view, the functional representations have, however, also a strong cognitive element. This is obvious in the set of functional representations that may be extracted from the user in relation to her information need (see the next section). The functionally different representations that may be generated from documents, e.g., title, abstract, introduction, headings, table captions, etc. of a scientific article, have a strong cognitive element too – see Figure 3.2 below. Because the rhetorical structure of scientific articles within a field has evolved in a continuous communication effort between active researchers over long time, these functional representations are socio-cognitive, and can be regarded as the distilled knowledge structures of a large number of actors. Indeed, investigations of the development of document types or genres (See, e.g., Swales, 1990) may be helpful in identifying representations with strong functional characteristics for use in IR.

Overlaps between *different* cognitive representations may be exploited because all the interactive communication activities in IR can be viewed as the result of cognitive processes in line with the cognitive viewpoint. Ingwersen formulates as a hypothesis (1996; 2001) that the more cognitively or typologically different representations (or evidence) that point to an information object – also over time – the higher the

probability that that object is relevant to the topic, the information need, the work task/interest situation at hand, or the socio-cognitive environment – as perceived by the information seeker(s).

The purpose of the theory 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 actor, as well as on combining representations from *different* cognitive actors in a structured framework. The two next sections examine the cognitive structures available in the cognitive space of the user and the information space of the system, followed by a section on the possibilities and advantages of combining them.

### *3.2.1 Polyrepresentation of the user's cognitive space*

Through a re-analysis of earlier investigations into the formation of the information need, Ingwersen arrives at a number of cognitive structures which form the basis for the application of polyrepresentation to the user's cognitive space (Ingwersen, 1996, p. 14-18). He sees the development of an information need as the result of communication, sensing or thinking processes, which result in the realisation that something is missing to solve a problematic situation. This conception is inspired by the work of, e.g., Mackey (Mackey, 1960), Dervin and Nilan (1986) and Belkin (Belkin, 1980; Belkin, Oddy and Brooks, 1982).

By discussing the stability of the user's mental structures in relation to the information need, he arrives at the four cognitive structures shown in the right side of Figure 3.1, and places the information need in relation to them in a causal structure: "It is the task/interest which, strongly influenced by the domain and dominated by the individual intentionality and cognitive state, causes a problematic situation and an information need to emerge." (Ingwersen, 1996, p. 15). The current cognitive state is "the little known about what is desired" (Ingwersen, 1996, p. 15). The current cognitive state, the problem space and the information need are all variable and easily affected by external input or thinking processes, while the work task/interest domain 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, or a person's continuing interest in certain aspects of Tolkien's artificial languages. 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.

Ingwersen connects this model of the four cognitive states to earlier empirical research and demonstrates that it can be inferred that these have been manifested in the studies, and that some of them appear in well-defined forms (1996, p. 16-18). Polyrepresentation of the user's cognitive space can thus be achieved by extracting a number of different functional representations from the user, as indicated in the middle of Figure 3.1. In an ideal situation, up to three potentially different functional representations may be extracted at any one point in time:

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 and domain description." (Ingwersen, 1996, p. 18)

Because the underlying cognitive structures are variable over a session, different versions of each representation may occur over time (this is indicated in the figure as an increasing number of 'boxes'). Ingwersen concedes that the extracted 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. These different types of information needs and their development are clearly demonstrated by the empirical studies (See, e.g., Ingwersen, 1982; Belkin, 1984), as is the role of the librarian in helping the user to define and refine her need.

The means by which the representations are to be extracted from the user in the theory of polyrepresentation are through a request model builder (RMB) interface. This should ideally be able to "…assess the proper nature of the information need and the underlying cognitive state…" and use the extracted representations accordingly, e.g., to assess if the user has a well-defined information need or not (Ingwersen, 1996, p. 21). Some of the prototypes reviewed in Ingwersen (1992, Chapter 7) had elements of this type of interface implemented by the aid of artificial intelligence architectures. In the view of the present author, the most serious challenge for proper polyrepresentation of the user's cognitive space is that very little research has been done, that can advice us as to whether such advanced request model builder interfaces can be made to function on a large scale. One of the problems with applications based on artificial intelligence is that

they require a large amount of domain knowledge to produce good results (Brooks, 1987; Croft and Thompson, 1987). A more modest approach without artificial intelligence features is planned for implementation in the TAPIR project. This project will attempt simply to extract different representations of the user's cognitive space and match these directly against selected representations from the information space. Because no end users could be involved in the main experiment of the dissertation, the users' information needs are represented by the static topics in the INEX test collection. The limited polyrepresentative qualities of these topics are discussed in Section 6.1.2 below, as well as the actual use made of the topics in the main experiment[7].

### *3.2.2 Polyrepresentation of the information space in IR systems*

The two major cognitive interactive structures in the information space of IR systems are the information objects and the IT components, see Figure 2.1. The possible representations one may generate from both are considered below, and related to the use made of them in the dissertation.

The *information objects* are influenced by several different cognitive structures. As described below the main experiment in the dissertation makes use of a corpus of scientific full text documents. This is a particularly rich source for generating representations and hence ideal for experiments involving the theory of polyrepresentation. Ingwersen regards the contents of the documents as reflecting the cognitive structures of the *author* "…in the form of signs, i.e., the transformations of the interpretations, ideas, and cognitive structures of the authors(s) with respect to their goals and intentionality." (2002, p. 289). Figure 3.2 below shows an overview of possible cognitively and functionally different representations of the information objects. In addition, the figure can be thought of as illustrating the sets of documents, and various overlaps between them, that may be retrieved in relation to a request using each of the representations (Ingwersen, 2002).

The functional representations originating from the author that may be extracted from the full text of the documents are shown on the right side of the figure. Many

[7] One should note, however, that the Title elements of INEX or TREC topic could be used to simulate a request, while the Narrative could simulate an extracted, corresponding problem description.

representations with strong functional characteristics are available because of the rhetorical structure of the scientific documents, as discussed above:

> "If we consider scientific communication by means of articles or monographs, the contents (and signs) are text structures, 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." (Ingwersen, 2002, p. 289-290).

Aside from the structure of the documents, the section titles at different levels, and the table and figure captions, represent functionally different ways of representing a document. These have previously been applied for document representation (Wormell, 1981). The use made of these representations in the main experiment is discussed in Section 6.2.2 below.

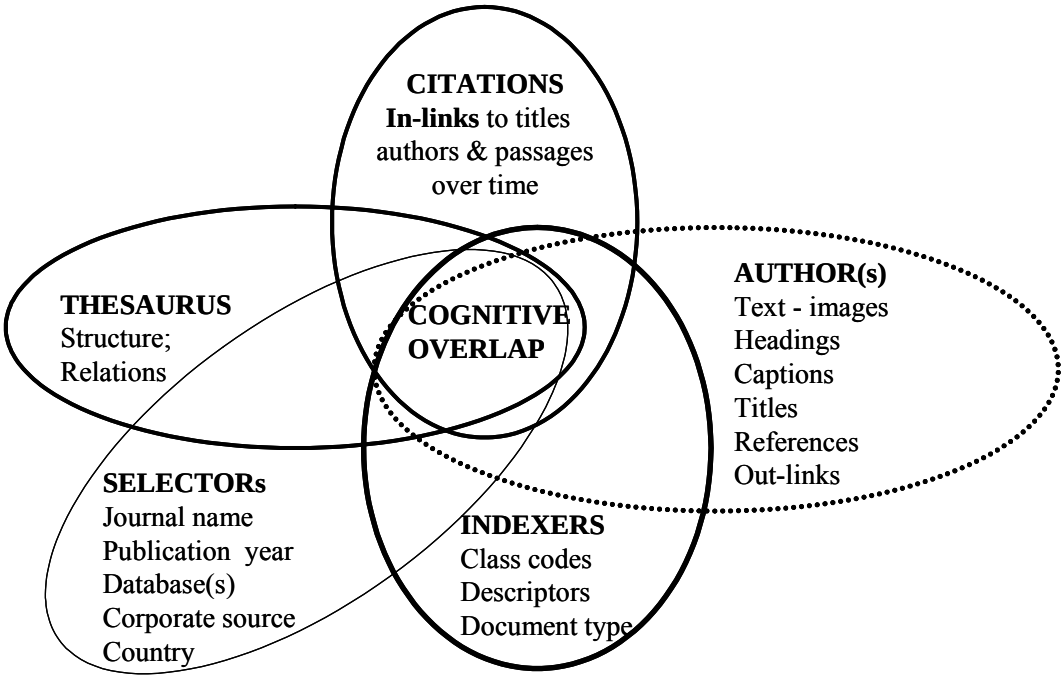



**Figure 3.2. Polyrepresentative overlaps of cognitively and functionally different representations of information objects. Retrieved sets are generated by one search engine and associated with one searcher statement. (From Ingwersen, 2002, Figure 1, p. 294; Extension of Ingwersen, 1996, Figure 5, p. 28).**

In addition, the *references* in the bibliographies may also be extracted from full text scientific documents. Ingwersen regards the references as representations of the authors' cognitive structures. The selection of particular references in a given document is, in Ingwersen's view, highly reflective of the situational factors that affect the author and her current cognitive state (Ingwersen, 2001, p. 295). Similarly, they may signal a kind of situational appropriateness to a potential user who, in her particular situation, might agree to the selection of references and thus find the document relevant. While the (citing) author is responsible for choosing *which* references to include, the cognitive structures of another agent, the cited author, play a strong role in the references as representation in the view of the present author. It is the cited author who has decided, e.g., the title of the cited document, and the characteristics of the cited document are (normally) outside the influence of the citing author, who can only refer to them as they are. As discussed below, other representations may be regarded as indirectly influenced by a variety of agents other than the author. However, the references given in the bibliography of a document (and the citations received by it) are very composite representations, and the direct result of complex interactions between several cognitive agents. As they are of special interest to the dissertation the characteristics of references and citations as representations are discussed in depth in Chapter 4. The actual use of references and citations in the dissertation is described in Chapter 5 and 6. An additional opportunity, offered by scientific documents in full text, is to identify the text surrounding the location in citing documents (if any) where the cited document is mentioned, and to use this as a representation of the cited document (See the top of Figure 3.2). This is proposed in Ingwersen (1996), but not explored further in the dissertation. The use of the anchor text of hyperlinks on web pages as a representation of the web page, which receives the link in web search engines, exploits the same idea (See, e.g., Brin and Page, 1998).

Aside from the author, the other major cognitive agent who typically produces representations of scientific documents is the human indexer: "Human indexers' cognitive structures are represented by the index terms added to the original information objects, these essentially being the result of an intellectual interpretation of an author's text or images, often guided by predefined rules and a thesaurus … for which other human beings are responsible." (Ingwersen, 1996, p. 8-9). The indexers typically select class codes and descriptors. These are taken from controlled vocabularies, e.g., a

thesaurus, which again is an interpretation of the vocabulary and semantic relations of concepts within the domain covered by the thesaurus. The thesaurus constructor therefore restricts the interpretation of the indexer and has an indirect influence on the representations made by the indexer. Indexer representations are available from the INSPEC database for the corpus used in the main experiment. Apart from the indexer representations shown in Figure 3.2 these also include uncontrolled terms and phrases (called identifiers in INSPEC). The identifiers represent the indexer's cognitive structures more directly, because they are chosen freely by the indexer. The thesaurus 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* introduced in Ingwersen (2002). These are shown in the lower left corner of Figure 3.2, and are an extension of the original figure from 1996. Ingwersen discusses these in relation to author, indexer and user aboutness:

> "In addition the model, [Figure 3.2], contains structures that are *selective* and different from those of indexers and users. Instead of aboutness, such features reflect *isness* by making available non-topical features connected to information objects—depending on media, domain, and presentation style. Most of the common bibliographic data or metadata thus belong to representations of isness. They are the result of selection or assessment processes performed by various actors on information objects and their authors over time." (Ingwersen, 2002, p. 293. Emphasis in original.).

Examples of these include the cognitive authority bestowed on articles or papers by the journal editors or conference committee when items are selected for publication in that journal (or conference). The publication year is also determined by the editor and not entirely by the author. Such actors are affected by their social/scientific context over time, Figure 2.1, the right-hand side. The authors' affiliations also possess selective power, e.g., by hiring particular researchers. No explicit use is made of these features in the dissertation, but future work might include, e.g., exploiting the impact factor of the journal as part of the weighting of representations[8].

[8] The 2002 journal impact factors of the journals used in the main experiment are listed in Appendix 2.

Most of the representations of the information objects are domain and media dependent. Although these factors clearly affect the results and execution of the experiments in the dissertation project, it is not known *how* because only one domain (computer science) is represented by the document corpus used in the dissertation. The exploration of differences to other corpora is very interesting and potentially promising for future work, but unfortunately it is outside the scope of this dissertation. Such future work might include studying the differences in characteristics between the corpus used in the dissertation and other corpora of scientific documents in full text, e.g., that offered freely by BioMed Central[9].

In Ingwersen's view, 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 essentially of interest to researchers in the system-driven tradition. A particular combination of these results in an implementation of a *retrieval engine* (for example the InQuery IR system used in the main experiment). From a cognitive viewpoint a retrieval engine is the embodiment of the ideas behind its construction. Ingwersen notes that the 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." (2002, p. 286). The ranking produced by one best match IR system is thus ultimately a cognitive representation of the knowledge structures of its designers[10], and hence of interest to the theory 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. This corresponds to the idea of data fusion in IR, as explored, e.g., by Belkin et al. (1995). Data fusion with several engines has not been tested in the present work, as only one engine (InQuery) was available.

---

[9] Approximately 3500 full text scientific articles formatted in SGML are available for download without charge from http://www.biomedcentral.com

[10] This is particularly obvious with the three IR models mentioned in section 2.1 above: Salton's Vector Space model is implemented in the SMART system, the Probabilistic Model is implemented in the Okapi system, and the Inference Network Model is implemented in the InQuery system.

In a sense, polyrepresentation of the user's cognitive space is less complex than polyrepresentation of the information space because there are many more cognitive actors involved in the latter (See Figure 2.1). However, the information space is more straightforward to handle because most of the transformations of the involved knowledge structures are manifested as tangible entities, e.g., in the form of published documents, assigned descriptors from existing thesauri, given references, implemented representation techniques and matching algorithms etc. Here the main challenge is to decide which representations to exploit and how to combine them with other representations. In the dissertation a selection of cognitive and functionally different representations is used (see Section 6.2 for details). Among such representations are document titles, abstracts, references, citations, and added descriptors, However, the experimental setting is such that the central object for investigation, the Boomerang Effect, is tested against two baselines: a baseline of pure polyrepresentative nature; and a traditional bag-of-words baseline. In this manner the project also attempts to address the value of the theory of polyrepresentation.

### *3.2.3 The polyrepresentation continuum*

In the following, the idea of a polyrepresentation continuum is proposed. This idea is inspired by experiences gained from working empirically with the theory of polyrepresentation, most notably the dissertation work and as supervisor of the MSc work done by Madsen and Pedersen (2003).

The holistic intentions of Ingwersen is obvious in any of his publications on the theory of polyrepresentation (e.g., Ingwersen, 1994; 1996; 2002). As such, theories, methods and results from both the system-driven tradition and the user-oriented tradition are amalgamated into a consistent and coherent framework that allows for both exact and best match principles to be combined in applications and implementations based on the theory. The theory is, however, inherently *Boolean* in much of its reasoning. This is apparent in the pronounced focus on cognitive retrieval overlaps, i.e., *sets* of retrieved documents based on different cognitive representations, see e.g., Figure 3.2 as well as the original figure 5 and 6 in Ingwersen (1996, p. 28 and 29). The appendix in the same article gives an example of cognitive online searching as an application of polyrepresentation, in which the Boolean derivation of a number of overlaps between two different search concepts or facets in three cognitive and functional representations

is demonstrated. The result of the strategy is a number of prioritised sets of documents, where the order of the sets is such that the first ones are constructed by the most intersections, i.e., the documents in them are in all or most of the overlaps between the representations. The following sets gradually involve fewer representations and finally fewer search concepts. In this way, the retrieved documents are ordered in a pseudo-ranking over the whole range of sets. A little discussed, but inherent point that may be learned from the appendix in Ingwersen (1996) is that the Boolean intersections, with the presence of all search concepts in all representations, ensure the *quality* of the sets that are matched. Without the structure created by the intersections and the resulting quality, it is doubtful whether the theory of polyrepresentation would provide the hypothesised improvements in performance. If, for instance, all search concepts were not present in each of the document sets generating the overlaps, the overlaps themselves would be of correspondingly lower quality.

For any implementations of the theory of polyrepresentation based on exact match the consequence is that a large and complex, but consistent set of overlaps have to be identified. This may be difficult to handle manually, but it can be automated without problems, and the quality of the set that the cognitive retrieval overlaps are based on can be maintained. If the implementation involves best match principles the situation is different. 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 theory of polyrepresentation, there is a risk that the quality of the sets that the cognitive retrieval overlaps are based on, as a whole, are too low. For instance, with two search concepts there is the risk that only the first of them is retrieved by some of the lower ranking documents in one representation, and the second in the lower ranks in another representation. Thereby proper polyrepresentation in the true sense of the concept cannot be achieved.

This problem was experienced by Madsen and Pedersen (2003) who tested the theory of polyrepresentation on the cystic fibrosis test collection (Shaw, Wood and Tibbo, 1991). Their initial implementation identified overlaps between cognitive and functional representations based on the requests in natural language in a best match environment. The resulting performance was very low, and an analysis revealed that proper polyrepresentation did not occur because only some of the search concepts were present in the documents in the overlaps, as described above. Madsen and Pedersen subsequently improved performance considerably by taking measures to improve the quality of the input sets. These measures included identifying the main facets in each request and performing manual query expansion which was adapted to each representation. The query expansion was necessary because the Boolean operations used to identify the overlaps between representations meant that the overlaps were empty in many cases without the expansion. The main conclusions of the study were that an implementation of the theory of polyrepresentation in a best match system has to take the Boolean nature of the theory into account by structuring the queries, and that, when this is done, the hypotheses in the theory can generally be confirmed.

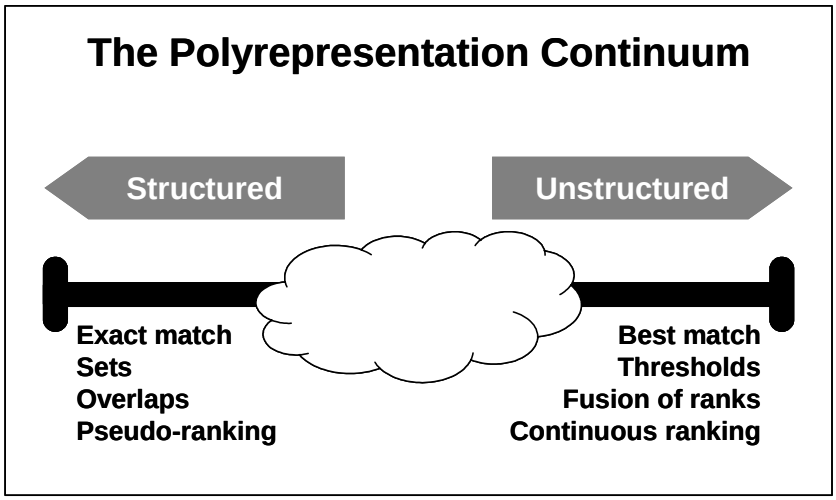



**Figure 3.3. The polyrepresentation continuum. Inspired by Madsen and Pedersen (2003).**

It is evident from the discussion above that *structure* is an integral part of the theory of polyrepresentation. The goal of the theory of polyrepresentation is to integrate this into a best match environment: "The core of the theory is, however, to explore the potential value of matching the multidimensional *cognitive variety* of representations inherently existing, extracted, or interpreted from information objects *and* from the cognitive space

of a user in a *best match* retrieval environment." (Ingwersen, 2002, p. 286-287. Emphasis in original.). The idea of a *polyrepresentation continuum* is therefore proposed in Figure 3.3 as a model for discussing how structured a given implementation of polyrepresentation is. At the *structured* pole of the continuum the implementations are based on exact match principles, leading to sets of retrieved documents for each representation from which overlaps can be formed and a pseudo-ranking be constructed. At the *unstructured* pole of the continuum the implementations are based on best match principles leading to a rank of the documents that are retrieved as input for polyrepresentation. Rather than generating overlaps between sets, the implementations at the unstructured pole 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 going from highly structured implementations to highly unstructured implementations.

The two implementations of the theory of polyrepresentation explored in the dissertation are close to either pole of the continuum: A fairly structured version of the boomerang effect was tested in the pre-experiment, and the best match boomerang effect tested in the main experiment was at the unstructured end. In relation to empirical investigations of the theory of polyrepresentation the middle of the continuum remains largely unexplored (hence the cloud), and could be the subject of future research. Softer versions of the Boolean operators, e.g., those available in the InQuery system make it possible to impose more or less structure on the queries sent to the system, as demonstrated for example by the weak and strong query structures examined by Kekäläinen (1999). Such query structures could prove favourable to future research with the theory of polyrepresentation, as indicated by the work of Madsen and Pedersen.

## 3.3 Summary statements

The theory of polyrepresentation has been presented in this chapter as a part of the cognitive viewpoint in Information Science. The strength of the theory is that it is *testable*, and that it points at many possible scenarios in which such tests could be carried out. In spite of this, the theory is still rather speculative in the sense that very few empirical investigations have been based directly on it. Borlund's analytically and empirically grounded evaluation package for interactive IR systems allows the user to develop her own information needs in relation to the simulated work task situations,

thus making these available in IR evaluation while at the same time providing experimental control (2000a). This is clearly in line with the theory of polyrepresentation on the user side, and would be the preferred methodology for evaluating IR systems based on the theory. No such implementations have, however, appeared yet. Part of the reason for this is probably the holistic intentions of the theory: it is a major research effort to establish both IR techniques in the information space *and* techniques to extract different representations of the user's information need within the same project. The present dissertation as well as the work of Madsen and Pedersen (2003) represents attempts to explore the potentials of polyrepresentation in the information space which in future research may be connected interactively to the user's cognitive space.

# 4 References and citations in IR

In her key note address at the European Conference on IR Research (ECIR) Spärck Jones briefly considered the Science Citation Index® (SCI) together with other automated IR services for scientific documents that were made possible from the 60s onwards by the aid of computers (Spärck Jones, 2003). She commented that the SCI was exotic as retrieval apparatus and that it remained on its own and never really related to the fully automatic statistical indexing and retrieval systems that were first being developed at the time. A few scattered studies from the system-driven tradition (e.g., Salton, 1971) and from the user-oriented tradition (e.g., McCain, 1989; Pao, 1993) have attempted to assess the usefulness of references and citations in IR. Generally, the results have been promising. In spite of these results and in spite of the widespread use and commercial success of the SCI the utilisation of references and citations as an integrated part of best match IR systems was not realised on a large scale until the construction of CiteSeer in the late 90s (Giles, Bollacker and Lawrence, 1998).

The aim of this chapter is to analyse the characteristics of references and citations as alternative representations of scientific documents in best match IR, and to review how they have been employed previously in both operational and experimental IR settings. The aim is to identify factors that might affect the behaviour and performance of automated indexing and retrieval techniques when references and citations are an integrated part of these techniques. This serves as a background for the proposed incorporation of references and citations into best match IR presented in chapter 5. *References* have probably played a role in individuals' search for scientific information ever since authors first began to refer to earlier research. A well-known search strategy is to consult the bibliography of a paper to identify earlier publications of interest from the references given (see e.g., Bates, 1979). Searching by *citations*, i.e., identifying publications of interest that are characterised by including particular references in their bibliography, requires, on the other hand, some sort of citation index such as those created by Eugene Garfield. Section 4.2 below gives a brief account of the motivation behind their creation, and the principal search strategies they offer as a background to the analysis in Section 4.3 of references and citations as alternative representations in IR. Section 4.4 reviews the results of earlier experiments with references and citations

in IR, followed by a description of a number of interesting proposals and applications, hitherto not tested for the performance. Section 4.5 contains a summary and discussion of the points covered in the chapter. As the terminology concerning references and citations is not entirely consistent in the literature the chapter begins with a brief terminological discussion.

## 4.1 Citations versus references

In the literature the term *citation* is often used both to denote the references in the bibliographies of individual documents as well as the citations received by a document. This may not seem unreasonable at first glance, but for most purposes it may be an advantage to distinguish between the two terms. The distinction made in the present dissertation is inspired by Price[11]:

> ”It seems to me a great pity to waste a good technical term by using the words *citation* and *reference* interchangeably. I, therefore, propose and adopt the convention that if Paper R contains a bibliographical footnote using and describing Paper C, then R contains a reference to C, and C has a citation from R. The number of references a paper has is measured by the number of items in its bibliography as endnotes and footnotes, etc., while the number of citations a paper has is found by looking it up in some sort of citation index and seeing how many other papers mention it.” (Price, 1970, p. 7, emphasis in original)

An important point about citations then is that they *do not exist* as such until the bibliographies of a number of documents have been indexed in a citation index. One may say that the references are transformed into citations when added to a citation index by a process of inverting them so that they are ordered, not by the referring documents, but by the documents that receive the citations. Adopting Price’s distinction provides greater terminological clarity, and may facilitate more substantiated analysis and discussion by making it apparent that references and citations do not necessarily exhibit the same characteristics. With Price’s distinction Wouters argues that “The citation is … a new sign with different semiotic properties from the reference.” (1999, p. 562).

[11] Salton (Salton, 1963) makes a similar distinction between a ‘reference index’ and a ‘citation index’, but does not define the terms reference and citation.

Following this argument references and citations are treated separately in Section 4.3 below.

A strict adoption of the distinction on a terminological level is hindered because there is no obvious broader term that covers both meanings. The field of webometrics, where relations of a similar type between documents are analysed, does not share this problem. Here the term *link* may be used as a broader term for out-links and in-links, respectively corresponding to references and citations (Björneborn and Ingwersen, 2001). Much of the terminological confusion in connection with references and citations is due to the convention of using the term citation in both meanings. For instance, much of the discussion on whether or not a theory of citation is needed (see for example *Scientometrics*, 1998, vol. 43, issue 1) has actually been concerned mainly with references and authors' possible motives for selecting particular ones for their bibliographies. Conversely, when employed, the term *reference* seems to be used fairly consistently in line with Price. In the dissertation Price's distinction is followed to the extent it is possible. A new broader term for references and citations does not seem readily available. The term 'link' will not be used in the dissertation as broader term in order not to draw too close parallels between hyperlinks and references and citations. Although they may be seem very similar concepts there are indications that motivations behind making a link on a web page are very different from the motivations behind selecting a reference to a scientific article (Thellwall, 2003)[12].

Rather than inventing a new term, the term *citation* will be used both in this function and in Price's sense depending on the context, much as in previous literature. Figure 4.1 below displays a network of documents and serves as an example of the relations and document roles that may need to be identified in analyses involving references and citations. Some of the documents may have several roles depending on the context of the analysis. In the network in Figure 4.1 there are four documents, three of which are indexed in a citation index as source items. The network contains the following relations and document roles:

[12] See also Fisher and Everson (2003) who found large performance difference in favour of a corpus of scientific documents versus a corpus of web pages when exploiting links and citations for text classification.

1) Relations

   a) **References** given to other documents (contained in referring documents R1, R2 and C2)
   b) **Citations** received (document C1 receives three citations, document C2 one)
   c) **Bibliographic coupling** (Kessler, 1965) between documents (R1 and R2 are bibliographically coupled as they share one reference to C1, thus having a coupling strength of one)
   d) **Co-citation** (Small, 1973; Marshakova, 1973) between documents (C1 and C2 are co-cited as R2 refers to both of them)

2) Document roles

   a) **Referring documents** (R1, R2 and C2 contain references to other documents)
   b) **Cited documents** (C1 and C2 receive citations; three and one respectively - R1 and R2 have not received any citations so far. Note that C2 is indexed in the citation index, whereas C1 is not)

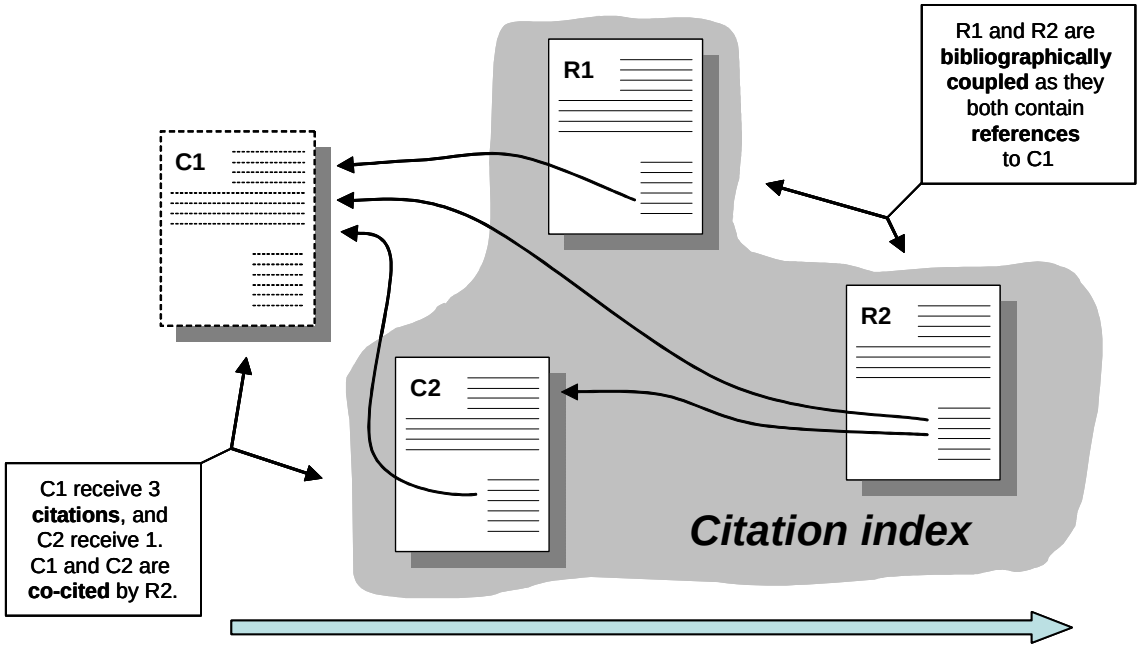



**Figure 4.1: Network of referring and cited documents illustrating the difference between references and citations, as well as the phenomena of bibliographic coupling and co-citation. Inspired by Price (1970), and Mählck and Persson (2000).**

As R1, R2 and C2 are indexed in the citation index as source items we may know a wide range of their characteristics, e.g., their bibliographical data including all author names, addresses, full title, journal names etc. or even the full text. Because C1 is not a source item in the citation index (indicated by the dashed lines in C1 in Figure 4.1) we do not know any characteristics of C1 other than what can be found in the references of the referring papers and what may be deduced from the roles and relations in the network. However, because cited documents may receive a considerable amount of citations, regardless of whether they are indexed in the citation index or not, the roles and relations in the network provide a rich source of data that may be exploited for a variety of purposes.

As a number of terms and concepts have become established in the literature they will be used in their original form here. These include:

- citer motivations (the motivations behind the inclusion of particular references by authors of referring documents)
- citation behaviour (the behaviour that may be observed from the bibliographies of published scientific documents)
- citation search strategy (any search strategy involving references and citations)

## 4.2 Eugene Garfield's citation indexes for scientific literature

While there are precedents like *Shepard's® Citations*[13] it was the citations indexes for the sciences created by Eugene Garfield that introduced citation indexing to a wider audience. Garfield presented the idea of a citation index for science as "a new approach to subject control" and describes it as "an association-of-ideas index" (Garfield, 1955). The main purpose of the index was the same as the existing discipline-oriented indexing and abstracting services of the time: to create a searchable index of scientific literature that may be used by scientists to retrieve documents on some subject. Compared to earlier indexes the major difference was that documents would be indexed not by

[13] Begun in 1873 Shepard's® registers US court cases to produce a citation index of the cases that have referred to each case subsequently. Shepard's® is now owned by LexisNexis™ and can be accessed at http://www.lexisnexis.com/shepards/.

subject headings assigned by domain experts, but by the references made by the authors of the documents. These references would then be inverted so that "…[i]n effect, the system would provide a complete listing, for the publications covered, of all the original articles that had referred to the article in question." (1955, p. 109). Although explicitly inspired by Shepard's®, the application of the idea of citation indexing to scientific literature is an original contribution, and represents a new principle for indexing and searching the scientific literature.

The primary aim of the proposal for a citation index for science "…was to improve the retrieval of scientific information" (Garfield, 1998, p. 68). The improvement is due to two factors; an economical one and the principle of citation indexing itself. The economical incentive behind a citation index is that each document does not need the subject analysis by expensive domain experts – instead "…although a great volume of material is to be covered, relatively unskilled persons can perform the necessary coding and filing." (Garfield, 1955, p. 111). Therefore, the compilation of a citation index would be, if not less expensive, then probably no more expensive than a conventional index, and more importantly, a citation index could provide better timeliness and greater coverage. The latter was experienced to be a considerable problem at the time as conventional indexes could have substantial backlogs, while their coverage was limited. The fundamental improvements of citation indexing over conventional indexes foreseen by Garfield in 1955 were:

1. The inherent terminological problems, e.g., changes over time and different uses across different fields, could to a certain extent be resolved. By referring to a previous work an author is in a sense interpreting the terminology, as well as "…indexing that work from his point of view." (1955, p. 110.) An added benefit is that a citation index does not require the scientist to learn an alien indexing language, but only to apply knowledge of existing articles that might have been referred to.
2. The equally inherent problems of relatively low specificity and exhaustivity experienced in practice with conventional indexes could be alleviated by using the references as index terms. Hence, the documents would be indexed by the micro unit of thought represented by the cited documents, and not by a few, more general subject headings. In this way a citation index could collate documents that would not have been brought together by conventional indexes.

3. By its very nature a citation index can be very helpful in "…tracking down the origins of an idea…" (1955, p. 110), and to assess the significance and subsequent impact of a publication. Works receiving many citations or with an interdisciplinary impact might be of special interest, but are usually very hard to identify using conventional indexes.

After a number of pilot tests the Science Citation Index® was published in 1963 in a printed version by the Institute for Scientific Information (ISI), a company founded by Garfield (Garfield and Sher, 1963). The index was multi-disciplinary, covered articles from 613 science journals published in 1961, and contained 1.4 million references (Weinstock, 1971). ISI has continued the production of SCI to this day, and has launched a range of services based on its citation database, among them the Social Science Citation Index (SSCI) and the Arts & Humanities Citation Index (A&HCI). At present the three citation indexes[14] cover more than 8900 journals adding approximately 1.1 million records per year containing 23 million references (Institute for Scientific Information, 2002). An interesting point about Garfield's initial proposal (1955) is that documents not indexed as source items were not intended to be included in the listings of cited documents either. By the time when the actual production of SCI began this had been changed and all cited documents were included, greatly enhancing the potentials of the ISI's citation indexes. The citation code developed by ISI to represent a cited document consists of a number of elements collected in the so-called *cited reference* (CR) string. The CR string is a reduction of the various forms in which citations to Garfield (1955) appear in the bibliographies of referring documents:

```
CR=GARFIELD E, 1955, V122, P108, SCIENCE
```

[14] All three databases are now integrated in the internet-based service *Web of Science* (for information see http://www.isinet.com/isi/products/citation/wos/index.html).

The different elements of the sting can be made searchable. For instance the online host Dialog[15] generates 4 indexes from the string, but does not make the volume number (`V122`) or the begin page number (`P108`) searchable:

| | |
|---|---|
| Cited Reference: | `CR=GARFIELD E, 1955, V122, P108, SCIENCE` |
| Cited Author: | `CA=GARFIELD E` |
| Cited Year: | `CY=1955` |
| Cited Work: | `CW=SCIENCE` |

### *4.2.1 Citation search strategies*

When used for IR purposes a citation index allows for a range of search strategies, some of which could not have been carried out in practice before its compilation. The main innovative feature of the citation index is that it facilitates the retrieval of unknown documents which refer to a given document of relevance. Garfield has described the search strategies that can be applied in SCI and their combinations with traditional strategies continuously over the years (e.g., 1964; 1970; 1979). Cawkell, a UK consultant for ISI, gives examples of a range of search strategies that are possible in the printed version of the SCI (1968; 1974), and later in the online versions (e.g., 1998; 2000). Citation searching may involve references, citations or both. In the following list the main types of search strategies possible in a citation index are summarised and generalised from the publications of Garfield and Cawkell:

a) *Backward chaining*, when references in the bibliography of a known document are retrieved. Although possible without a citation index this strategy may be greatly accelerated by a citation index.
b) *Forward chaining*[16], where documents that refer to a known document of relevance[17] are retrieved (usually the seed document must be a few years old in order to have received citations). This is the main innovation offered by a

---

[15] http://www.dialog.com

[16] The terms backward and forward chaining as used here were coined by Ellis (1989).

[17] These known documents of relevance are often called *seed documents*, see e.g., Pao and Worthen (1989). They need only to consist of a *reference* to an actual document – the actual document itself in physical form does not need to be at hand. Although they are most often called seed documents in the literature they might simply be called *seeds*.

citation index as mentioned above, and it is a core component of strategy c, d, e, f and h below. Note that the seed document itself is not necessarily indexed in the citation index. It is sufficient that other documents in the citation index have referred to the seed document to make forward chaining possible (See Figure 4.1).

c) *Citation cycling*, where the starting point is a *backward chaining* from the references of known (recent) documents, which are then used as basis for a new *forward chaining*. This cycling activity can continue in several iterations.
d) *Uncontrolled subject search*, where unknown documents are retrieved by words from titles (and today also from abstracts and author keywords) and used as the basis of a cycling search.
e) *Controlled subject search,* where unknown documents are retrieved using descriptors etc. from a domain specific index or abstracting journal as the starting point for a cycling search.
f) *Highly cited document search,* where unknown documents with a high number of citations are retrieved based on a specific starting point, e.g., highly cited documents by a known author, or documents that are highly cited within a specific subfield.

To this may be added strategies based on bibliographical coupling (Kessler, 1963), and on co-citation (Small, 1973):

*g) Bibliographical coupling search,* where documents are retrieved which are related to a known, relevant document by one or more bibliographical couplings.
*h) Co-citation search,* where documents are retrieved that refer to two (or more) known, relevant documents in their bibliographies.

In the printed versions of ISI's citation indexes the bibliographical coupling search (g) would be quite laborious if a coupling strength of more than one is desired. After initial experiments (Vladutz and Cook, 1984; Small, 1987) the strategy was automated in the CD-ROM version of the citation indexes as a 'related records' feature (Garfield, 1988), and is also now a part of the Web of Science. The related records feature retrieves documents that are bibliographically coupled (if any) to the document currently being examined, and ranks the result descending by the coupling strength. For online use of the strategy Christensen and Ingwersen (1996) demonstrate how bibliographic coupling can be performed at the online host Dialog. Although relatively straightforward to carry out online as demonstrated, e.g., by Chapman and Subramanyam (1981) co-citation

search (type h) does not seem to have received much attention for retrieval. Instead co-citation has been used extensively for mapping the structure of research fields (see e.g., Small and Griffith, 1974; Griffith et al., 1974; White and Griffith, 1981; White and McCain, 1998).

The initiation of the citation search strategies outlined above depends largely on human intellectual effort during the search process. The effectiveness of the seed document approach is heavily dependent on the relevance of the chosen document to the information need of the user, as noted, e.g., by McCain (1989). The cycling search strategy is equally dependent on the initial document. It may seem less so if the controlled or uncontrolled subject search strategies are employed as a starting point for the cycling. However, if several iterations are to be carried out, human judgements are needed to decide which documents to pursue further, and which to discard (see, e.g., Cawkell, 1968). All strategies will in practice retrieve an exponentially growing number of documents per iteration if such judgements are not made. This is also the case for citation cycling based on known documents. While it is impossible or at least hugely laborious to follow all references and citations in the printed version of ISI's citation indexes, it is not so in the electronic versions, e.g., Web of Science. Here all the source items are hyperlinked and thus it is quite fast to browse from one record to another. However, as with all search strategies based on browsing there is a risk that the user may get lost in the information space quite rapidly and become disorientated (Conklin, 1987; Rivlin, Botafogo and Shneiderman, 1994; Otter and Johnson, 2000). Very few studies have investigated how large a role citation search strategies play in scientists search behaviour. In a study of the information seeking activities of social scientists Ellis (1989) has shown that most of them use citation search strategies to some extent, and that they are thought of as being very fruitful. The most widely applied strategy is backward chaining, whereas forward chaining was not generally known (although it was used by almost all of those few who did.) Controlled experiments with comparative tests of the efficiency of citation search strategies versus conventional ones are also rare. The studies that incorporate such analyses in operational or experimental IR settings are reviewed below in Section 4.4.

## 4.3 References and citations as alternative representations

After the publication of the first edition of the SCI much debate arose concerning its value as search tool and especially its possible use or misuse in evaluation of scientists. This was partly due to the controversial nature of SCI's approach to bibliographical control, which required far less intellectual work than existing systems. The core of the criticism that was raised, and continues to be raised, concerns the authors' motives when including, or excluding, particular references in the bibliography of a scientific paper. As mentioned above Garfield himself presented the citation index as "an association-of-ideas index" in his 1955 *Science* article. The underlying assumption behind this is that there is some kind of semantic relationship between the referring and the cited document, tying them together. But apart from focussing on the citation index's advantage of providing means to assess the impact and significance of a publication through tracing subsequent references, he did not reflect on how the references themselves have been chosen, and the consequences for a citation index. Perhaps this is not unnatural as Garfield was writing as a scientist to fellow scientists who at the time might be expected to share a common perception of how to refer to earlier works. As later remarked by Merton in his foreword to Garfield's book on citation indexing:

> "It was of course unnecessary for Eugene Garfield to identify this composite communications-intellectual-property-and-reward system in order to arrive at his concept of the citation index. He needed only the sense that the system provided the ingredients for systematically identifying, through citation indexing, links between the works of scientists that could be put to use both for searching the literature and for exploring cognitive and social relationships in science." (1979, p. viii)

Indeed, Cronin sees the launch of the SCI as the actual birth of the widespread interest in authors' citation behaviour:

> "It was as if the scientific establishment had not previously recognised the full import of one of its most frequently exercised conventions — citation. The development of citation indexes for science turned the spotlight on a little-thought-about practise.…Attention was now being focussed on the 'why' as well as the 'what' of citation." (Cronin, 1984, p. 8)

Since then many opinions have been vented and much research conducted with the purpose to shed light on the phenomenon. The 'why' has been addressed by analyses of citer motivations, and the 'what' has been examined by studies of citation behaviour. The next sections will discuss the nature of the semantic relationship between referring and cited documents: Section 4.3.1 and 4.3.2 examine this from the perspective of references in Price's terminology, and Section 4.3.3 from the perspective of citations.

### *4.3.1 Citer motivations and citation behaviour*

A considerable amount of papers have been published on the subject of citer motivations and citation behaviour. It is important to note, however, that most of them are concerned with references in Price's terminology.

In his 1984 essay and review *The citation process – the role and significance of citations in scientific communication* Cronin finds that there is no all-embracing theory that can account satisfactorily for their characteristics. He identifies two theoretical positions in the literature from which work on the subject have been approached:

1. A normative 'storybook' image that understands science and the act of giving references as a "serious activity, governed by a tacit understanding of how and why authors should acknowledge others. That is to say: an author's reasons for citing in a particular way at a particular time are controlled by an internalized set of norms." (p. 2). This position is inspired, e.g., by the constructivist works of Robert K. Merton, and offers a coherent theory albeit with certain limitations.
2. A 'relativist' account (in opposition to the normative view) that focuses on science as a social process, and the actions and behaviour of scientists as context dependent. This may be characterised as a social constructivist position (Latour and Woolgar, 1979), which poses questions more than providing a coherent theory.

The normative position considers scientific publications including their references as an important part of the primary, formal communication system in science. Merton (1942) views this as an institutionalised system which, at the same time, is a communication system and a system for distributing rewards. Published works, and the recognition given to them in the form of citations, permits the individual scientist to accumulate "reputational wealth" through peer-recognition (Merton, 1988). The selection of references is seen as a rational, objective activity that is bound to the moral and ethical

norms and rules for good practice, which are adhered to voluntarily for the common good and benefit of all in the scientific community. First and foremost references reflect which earlier works that are being built upon. Hereby the references become a sort of payment for using others' work, and add to the reputational wealth of the cited. Garfield is clearly in line with this position, see e.g., the 15 reasons for why references are provided in scientific papers listed in Figure 4.2a (Garfield, 1965).

Research on citer motivations from the social constructivist position was initiated by Kaplan (1965). Kaplan pointed out that at the time, apart from anecdotal evidence, very little was actually known about the norms of citation behaviour operating in actual practice. He identifies the most important research question as "…the extent to which citation practices reflect significant elements of the normative and value systems of scientists." (1965, p. 183). Since then research from the social constructivist position has proposed many other factors than those offered by the normative view that may affect citer motivations. An example of such factors is Thorne's (1977) list of strategies that authors may apply to improve their chances of getting published, the majority of which relates to selecting references (these are summarised in Figure 4.2b). Gilbert agrees that bestowing credit and recognition are only secondary functions of references. He considers references as tools of persuasion that contribute to the "...demonstration of the validity and significance of the work reported in scientific papers." (Gilbert, 1977, p. 116). Similarly, Cozzens (1989) regards references as rhetorical devices. It is important to note that the social constructivist position does not dismiss that there are normative factors at play in giving references. It rather questions that they should be the only or primary ones. Based on a small study MacRoberts and MacRoberts (1986) argue that as little as 30 percent of the formal references that ought to have been cited according to the norms are actually cited.

| a) Garfield's 15 reasons for providing references in scientific papers | b) Summary of Thorne's strategies to improve authors' chances of getting published related to giving references |
|---|---|
| 1. Paying homage to pioneers<br>2. Giving credit for related work (homage to peers)<br>3. Identifying methodology, equipment, etc.<br>4. Providing background reading<br>5. Correcting one's own work<br>6. Correcting the work of others<br>7. Criticizing previous work<br>8. Substantiating claims<br>9. Alerting to forthcoming work<br>10. Providing leads to poorly disseminated, poorly indexed, or uncited work<br>11. Authenticating data and classes of fact—physical constants, etc.<br>12. Identifying original publications in which an idea or concept was discussed.<br>13. Identifying original publication or other work describing an eponymic concept or term as, e.g., Hodgkin's Disease, Pareto's Law, Friedel-Crafts Reaction, etc.<br>14. Disclaiming work or ideas of others (negative claims)<br>15. Disputing priority claims of others (negative homage) | 1. Hat-tipping citations<br>2. Over-detailed citations<br>3. Over-elaborate reporting<br>4. Evidentiary validity (references can be selected to support any desired point)<br>5. Self-serving citations<br>6. Deliberate premeditation (conscious playing of the citation game)<br>7. Citations as projective behaviour (citations as reflection of author biases)<br>8. Conspirational cross-referencing (citing each other's work collusively)<br>9. Pandering to pressures (citing works because it is felt that the reading public requires, or expects, them to be cited)<br>10. Intra-professional feuding<br>11. Obsolete citations (referring to work that has been proven invalid or useless)<br>12. Political consideration (citing by the 'party line') |

**Figure 4.2: Citer motivations as seen by a) Garfield (1965, p. 85), and b) Thorne (1977, p. 1159-1160) representing normative and social constructivist positions respectively.**

Empirical evidence tends to support elements of both positions. Small (1982) reviews studies of what he calls *context* and *content* analysis of references. The former studies the citation behaviour by examining the context of the references as they appear in the text. The investigations of Lipetz (1965) and Duncan *et al.* (1981) are interesting from an IR perspective: In these the main intention is to improve the effectiveness of citation indexes as IR tools by supplementing each reference with additional information concerning its *role* in the referring document. The idea was first put forward by Lipetz (1965), who proposed a range of relational indicators designed to aid a user during the search process by pointing to the function of the reference in a referring paper. Both

Lipetz and Duncan *et al*. (1981) succeed in assigning this type of operators on small test corpora. However, attempts to implement the idea on a large scale failed, essentially because the assignment was too resource demanding as the roles had to be assigned intellectually. The main economical incentive behind a citation index is thus lost. Cronin notes that most of the context analysis of references are of a pragmatic nature and that they implicitly share a normative view in that all their categories are of the 'serious' kind (1984, p. 41) such as those in Figure 4.2.a.

Not all attempts to discover the nature of the relationship between the referring and the cited document are based on intellectual effort though. Herlach (1978) proposes that documents, which refer multiple times to a given seed document in the full text, are more likely to be relevant than those that only refer once to the seed document. Such documents with multiple *mentions* of a seed document can be automatically identified from the full test without the need for intellectual effort. Herlach investigated this idea in a small study of six seed documents and the documents that refer to these in the areas of endocrinology and radiology (1978). She found that the documents with multiple mentions to the seed documents were more closely related to the seed documents topically, and also more useful as seen by human judges. However, the study also indicated that a relatively large proportion of the documents with only a single mention of the seed documents were also topically related and useful, although not to the same extent.

The studies involving *content analysis* of references are of a more sociological nature. Most of these studies attempt to analyse citer motivations with the purpose of studying fundamental aspects of these by examining the surrounding text. Examples include Moravcsik and Murugesan (1975) and Chubin and Moitra (1975) who study the context as well as the quality of references. By doing so they identify those references that they deem to be central to the research presented (i.e., in agreement with the motives put forward by the normative position), and those that may be considered to be included for other, more dubious, reasons. For instance Moravcsik and Murugesan (1975) found that as much as 41% of the references analysed in their study of theoretical high energy physics papers were of what they called the 'perfunctory' type; that is, they were references to several papers that more or less make the same point without contributing further to the referring document, included perhaps in the hope of making several people happy for rhetorical reasons.

As noted, e.g., by Gilbert (1977), a problem in both context and content analysis is the difficulty of unambiguously determining the correct function of a reference from its appearance in papers. In these attempts to classify references *post hoc* the underlying and inherent problem is, in my view, that the methods used are based on *inference*. As already realised by Chubin and Moitra (1975) the act of giving references is ultimately an internal and private cognitive process. Therefore, although shedding some light on the 'what' of giving references, all the attempts to classify references mentioned above are unlikely to provide any profound answers as to the underlying 'why'. Because authors' motives for giving references do not always adhere to the scientific norms the direct interviewing of scholars about their citer motivations such as carried out by Brooks (1985; 1986) have the same weakness. Cozzens points out that, although analytically distinct, the motives put forward by the two positions are concretely indistinguishable (1989, p. 440). The consequence is that analysts may discuss them separately, but they are both present in any act of giving references. Cronin concludes his essay laconically by stating that citer motivations cannot be said to adhere to a set of universally recognised norms, nor can they be said to be given completely randomly or inconsistently. In his opinion giving references is a highly complex phenomenon, which needs to be thought of as a process. Thus a given list of references reflects both the author's personality and the professional environment in which it is created.

### *4.3.2 References as 'concept symbols' in IR*

The potential value of *references* as alternative representations in IR is obviously closely related to the citer motivations advocated by the normative position. If references are regarded as conforming and consistent acknowledgments of intellectual debt providing "…formal, explicit linkages between papers that have particular points in common…" (Garfield, 1979, p. 1) it is clear that there is a reasonable probability of some sort of semantic relationship between the referring and the cited document. From this follows that by retrieving the documents referred to in the bibliography of a relevant document one may have a fair chance of identifying a good deal of relevant documents published previously. It is not so clear, however, what the effect on IR will be if citer motivations are determined to a great degree by the factors maintained by the social constructivist position. A marked tendency to omit certain references and to refer biased and unnecessarily frequently to colleagues, well-known peers in the field, citation classics, by political considerations etc. can be expected to decrease the

performance of citation search strategies. However, as pointed out by Kaplan (1965), it is exactly the extent to which references are given by the scientific norms that is the important issue. Although the norms and rules of a specific research field are perhaps not adhered to consistently, they cannot be transgressed repeatedly by an author if he wishes to publish his work. As indicated above the chosen references have to function effectively as tools of persuasion (Gilbert, 1977) and rhetorical devices (Cozzens, 1989). Thus, for IR purposes some of the non-conforming types of references may not *per se* lead to poor retrieval results. For instance the high percentage of perfunctory references identified by Moravcsik and Murugesan (1975) might not harm IR performance as they identify a group of papers making the same point.

In the view of Small (long time Director of Research and Chief Scientist at ISI) references can best be conceived as *concept symbols* which stand for an idea that is being used in the course of an argument (Small, 1978). This is a resonance of the normative position, but in Small's opinion it also includes social or political functions of references, because "[w]hether the motives for citing a work are politically conditioned or merely haphazard…the work must be associated with specific language in the text and cannot be appended without some explicit or implicit context." (1978, p. 337). Though this does not explain why particular authors select particular references to symbolise particular concepts (Morman, 1981, referred in Cronin, 1984) it does go some way towards bridging the gap between the normative and the social constructivist positions. From the point of employing references in IR the idea of references as concept symbols is attractive in my view: regardless of whether or not some references are omitted, forgotten, biased etc. the references *actually given* function as symbols for a concept.

As mentioned above in Section 3.2.2 all cognitive and functional representations can be expected to be domain dependent. References and citations is no exception, for instance, Hjørland (1993) notes that citation rates are likely to vary across domains, and Seglen has shown differences in citation rates between basic and clinical medical research fields (1997). Because only one corpus covering a particular domain (computer science)

is used in the dissertation it is not known how this affects the results[18]. If such differences could be shown and operationalised in corpora used for IR purposes, some of the normalisation techniques used in scientometric research evaluation (See, e.g., Moed, De Bruin and Van Leeuwen, 1995; Ingwersen, Noyons and Larsen, 2001) to level out domain differences might also be of value when utilising references and citations in IR.

Small's notion of references as concept symbols provides us with an explanation of the character of the semantic relationship between referring and cited document, albeit with some uncertainty as to why particular references have been chosen to represent particular concepts. As discussed in Chapter 3 this uncertainty is not necessarily a disadvantage seen from a cognitive viewpoint: the theory of polyrepresentation may be used to reduce and exploit the uncertainty to improve results. In the present citation indexes it is mainly up to the user to reduce the uncertainty. As discussed above in Section 4.2.1 this is done by making intellectual choices of which of the available references and citations to follow. The boomerang effect presented in Chapter 5 below is an attempt to use the theory of polyrepresentation to reduce the uncertainties and unpredictability associated with references and citations as representations. As discussed in Section 3.2.2 above the selection of references can be regarded as highly reflective of the situational factors that affect the author and her current cognitive state at the time of writing. From a cognitive viewpoint when reference lists in large numbers are transformed into citations by citation indexing the citations become expressions of socio-cognitive judgements with increasing statistical confidence as more citations are received. This is in line with the view on citations presented below.

### *4.3.3 Citations as a statistical phenomenon*

During the last decades much of the criticism against citation indexing has been raised not against its use as an IR tool, but rather against its use and misuse in citation analyses where the quantity of citations received by, e.g., an author is regarded as an indicator of research quality. It is not unfair to say that it is the proliferation of citation analyses that began shortly after the publishing of the first edition of SCI that has mainly fuelled the

[18] The journal impact factors for the journals corpus used in the main experiment are listed in Appendix 2, as an indicator of the differences within the corpus.

discussion on citer motivations and citation behaviour. From the social constructivist point of view some quantity of the references in any specific paper will inevitably by given for reasons that do not conform to the scientific norms. This has repeatedly been given as an argument against citation analyses of research quality. For instance, MacRoberts and MacRoberts dismiss citation analysis as a valid method entirely:

> "When only a fraction of influences are cited, when what is cited is a biased sample of what is used, when influences from the informal level of scientific communication are excluded, when citations are not all the same type, and so on, the "signal" may be repetitive, but it is also weak, distorted, fragmented, incoherent, filtered, and noisy." (MacRoberts and MacRoberts, 1989, p. 347)

An important question therefore is whether this critique bears on *citations* when they are used as representations in IR.

The main counter argument put forward by proponents of citation analysis is a statistical one, which is rooted in the distinction between references and citations such as formulated by Price (1970). It takes as a starting point that citations have a different set of characteristics from references, as citations are aggregated from many different sources in a citation index. In a paper explicitly written to refute social constructivist criticism of citation analysis van Raan states:

> "Indeed, if just … one paper with its peculiar references would be analysed, a seriously mistaken picture of the field concerned will be obtained. But as soon as further papers are added, similar but also other irregularities will be discovered in their reference lists. Does this mean that one would never be able to get any sensible idea of the most important work in that field? This is statistically only the case if *all* researchers refer to earlier work *completely arbitrarily*…. As soon as authors refer, already to a small extent 'reasonably', i.e., not based on a 100%-random 'reference generator', valid patterns in citations will be detected if a sufficiently large number of papers is used for analysis. Furthermore, it is statistically very *improbable* that *all* researchers in a field share the same distinct reference-biases (for instance, *all* authors cite *deliberately* earlier papers which did *not* contributed whatsoever to their field) …. Even *if* authors would cite all relevant work, this would be statistically unnecessary, as 'incomplete' reference lists will already provide…significant patterns." (van Raan, 1998, p. 134-135. Emphasis in original.)

This view on citations, which be may called a statistical view, is shared by a number of the researchers engaged in citation analysis (e.g., Martino, 1971; Small, 1978; Cole and Cole, 1987; Wouters, 1999). When applied to research evaluation the view rests on two statistical factors.

The first factor concerns the question of whether or not "…a sufficiently large number of papers is used for analysis." (van Raan, 1998, p. 134), i.e., if enough *cited* documents are being used as source material. In research evaluation this can be ensured by, e.g., excluding a department from an inter-departmental comparison[19]. The underlying idea behind the formulation that "a sufficiently large number of papers" must be used for a citation analysis is that the unit under examination (e.g., a researcher, a department, a country etc.) will not be *characterised sufficiently* by the papers used to represent the unit if this is violated. An actual minimum number of papers to use is rarely given and there exists few guidelines on how to set such a threshold.

The second factor is related to the (often not explicitly stated) fact that the citation indexes from ISI are multi-disciplinary and index a very large number of journals. The sheer number and variety of source journals giving citations means that giving prominence to highly cited units (authors, departments, countries etc.) in a citation analysis will, with high certainty, also separate the influential units from the non-influential ones. Based on the binominal probability distribution Cole and Cole (1987), for instance, calculate that it is very unlikely that an influential author will not receive a markedly higher number of citations compared to other authors in a field. In support of his arguments van Raan (1998) points to a number of studies which demonstrate that citation counts correlate with quality indicators based on peer review. The statistical view lies behind the numerous bibliometric studies of the structure of research fields, the relative performance of institutions and countries, etc. Note that in many of these,

[19] This can be done, for instance, if the department's share of publications indexed in the citation indexes is very low compared to the other departments analysed. This is an indication that the department's field of research is not well represented in the citation indexes. Therefore an evaluation based on these might not produce valid results. For example Ingwersen, Wormell and Larsen (2000) exclude 4 of 13 research centres from their citation analysis because the 4 centres displayed noticeably different publication profiles from the rest.

higher level aggregates are being analysed, e.g., the number of citations received by the documents published by an institution or whole countries.

In IR, *citations* may have two major roles (in Price's terminology): they may either be used as seeds in a forward chaining, or their frequency may be utilised in some way similar to the highly cited document search (type f mentioned in Section 4.2.1 above[20]). In both roles the cited documents function as a representation of the information need, and they are the search keys that are matched with the index keys from the documents. The way that two factors of the statistical view will influence the IR results depends on the role that citations play in the IR method. In Figure 4.3 we map the consequences of the two factors against the role of citations in IR as an aid to discussing them.

When used as seeds in a forward chaining (Role A, Figure 4.3), the set of referring documents that give citations to the seed document will be retrieved. One might argue that seed documents function as references, but the common characteristic that defines the set of retrieved documents is that they all give a citation to the seed document in question. The size of the citation index (the second factor, Figure 4.3) may be of minor importance in forward chaining. Given a set of seed documents that can characterise the information need sufficiently there is a good possibility that a large range and variety of referring documents will be retrieved when relying on ISI's citation databases. If a smaller custom citation index is used fewer documents will probably be retrieved. Therefore, in terms of absolute recall, it is obviously an advantage to use ISI's citation indexes. In both cases, however, there is a good chance of achieving high precision so that a fair proportion of the retrieved documents will be relevant, as indicated by the studies reviewed in Section 4.4 below.

[20] If search strategies based on co-citation were implemented, citations would be the main component in them.

| | **Role of citations in IR** | |
|---|---|---|
| **Statistical factors:** | *A. Seed documents in a forward chaining* | *B. Retrieval of highly cited documents* |
| *1. Sufficient characterisation of the information need* | Selecting good seed documents is critical for the success of a forward chaining | The starting point, which determines the frame within which the highly cited documents are identified, is critical |
| *2. Size of the citation index* | Size is of minor importance: a small citation index will result in fewer documents retrieved, but precision is not necessarily harmed | A large citation index will give better evidence of citation frequencies, and better protection against irregularities |

**Figure 4.3: Juxtaposition of the two factors of the statistical view on citations against the two roles of citations in IR.**

Characterising the information need sufficiently (the first factor, Figure 4.3) is far more critical in a forward chaining (Role A, Figure 4.3): if the seed documents used for a forward chaining do not represent the information need well both recall and precision will suffer as indicated by McCain (1989). When exploiting citation frequencies to retrieve documents because they are highly cited[21] (Role B), the size of the citation index (the second factor) is a great deal more important: Clearer evidence of the difference in the citation frequencies between particular documents will be available in a large citation index with a much greater number of source items. Exploiting citation frequencies from a small citation index may result in bad retrieval performance as there is a greater risk that individual biases in citing behaviour may have a greater influence on which documents that will be retrieved. The selection of a good starting point for the search is critical because it determines the frame within which the highly cited

[21] An example of such a search strategy could be: Retrieve all documents on a topics from the last few years in a citation index using words from the title and abstract, extract the references from the retrieved documents, and retrieve the top(n) cited documents as the result of the strategy. This strategy can be carried out at the online host Dialog.

documents are identified. Hence a sufficient characterisation of the information need (the first factor) is as important as in a forward chaining.

## 4.4 References and citations in IR R&D

Several researchers within IR have been interested in assessing the usefulness of references and citations for a variety of purposes in their research and development (R&D). The studies that report the results of these efforts are scattered in the literature from the two main research traditions in IR: researchers from the system-driven tradition have mainly experimented with citation data as part of document representation in laboratory settings, while researchers from the user-oriented tradition have tested the pragmatic use of citation databases in operational settings with the involvement of users. In addition a number of interesting uses have been made of references and citations in untested prototypes or operational systems. The following three sections review these studies and systems.

### *4.4.1 Experiments in laboratory settings*

It is interesting to note that some of the early test collections made use of citation data. The requests used in the Cranfield II experiments (Cleverdon, Mills and Keen, 1966) were constructed by asking authors of selected published documents (referred to as base documents) what initial question led to the writing of the paper. Part of the documents that were subsequently selected for relevance assessments were documents referred to in the base documents as well as documents bibliographically coupled to them (Ellis, 1996). Another example is the CACM test collection (Fox, Nunn and Lee, 1988) which contains information on direct citations, bibliographic couplings and co-citations among the documents in the collection. Generally the studies made from the system-oriented tradition focus on the ability of representations generated from citation data to represent the content and subject matter of documents much in the same way as representations generated from terms.

The earliest formal study of the possible benefits of citations in experimental IR appears to have been carried out by Salton. In a small study the similarities of manually assigned index terms are compared with representations made from references and citations in 62 documents (Salton, 1962; 1963). The end goal is to use the results in the

development of methods for assigning terms to the representation of a document from other related documents. The results indicate that there are good correlations between the manually assigned index terms and the representations made from both references and citations. It is cautiously concluded that if other studies could confirm the results "…citations could provide a large number of relevant index terms not originally available with a given document collection…" (Salton, 1963, p. 456-457). Following the definition of bibliographic coupling (Kessler, 1963) Kessler compared how 334 articles from *Physical Review* were grouped by an analytic subject index to the groups formed by the same papers by bibliographic coupling(1965). He found high correlations between the two groupings, and that a higher coupling strength results in a higher probability that two papers were related by the subject headings. No actual experimentation was done to determine the effect on IR performance in these initial studies.

A more straightforward approach to the utilisation of citation data in IR is taken in an experiment carried out by Salton (1971). Here the idea is not to use the citation data for adding terms from associated documents as proposed in Salton (1962; 1963), but to test their usefulness directly as identifiers of document content. A subset of 200 documents and 42 queries from the Cranfield test collection (Cleverdon, Mills and Keen, 1966) was used in the experiment. Citation-based document representations were constructed by adding codes representing the documents' references to the standard term-based document representations in the SMART system. In order to match the citation codes assigned to the documents with the queries, some way also had to be found of expressing the information need in terms of citation codes. Salton solved this conveniently by generating citation representations from the base document for each query in the Cranfield collection. This was done by adding the references found in the base documents to the query vector, thereby in effect employing them as seed documents in a forward chaining. The result of a match between a query and a document representation in the new citation-based approach was documents that were bibliographically coupled to the base document. In addition a citation code representing the document itself was added to the representation of each document so that it would be retrieved if a reference was made to it in a base document. Being aware that the base documents and references therein were used to form part of the recall base in the Cranfield collection, measures were taken by Salton in the experiment and the interpretation of the results to avoid bias in favour of the citation-based approach. Tests

were run on each representation type separately, and on combinations of them. The results indicate that:

1. citation data, when available, are generally useful as content indicators;
2. that on their own citation data are fully comparable in retrieval effectiveness to standard term-based representations at the high precision–low recall end;
3. that the combination of standard term-based representations with citation-based ones provide considerably better retrieval performance compared to term-based representations alone. (Salton, 1971, p. 109)

Salton notes that seed documents should be taken from documents whose strong relevance characteristics to the query are known by the user. For practical implementations of the approach Salton recommends that information on good seed documents are collected from users at query time. In my view the main weakness of Salton's experiment is the use of base documents for the generation of the citation-based query representation. The particular method of construction in the Cranfield collection is so closely related to the generation of the citation-based query representations in Salton's experiment, that the possibility of biases in favour of the citation-based approach cannot be ruled out, even with the efforts taken to avoid it in the experiment. As with many of the early IR experiments the result need to be interpreted with caution because of the limited size of the test collections used. Nevertheless the study is significant because it is the first of its kind that formally examines the effect of citations as *direct* identifiers of document content in an IR experiment. A further strength of the experiment is that it demonstrates that citation data can be incorporated in a best match system on equal terms with term-based representations. The important question quite naturally not investigated in the experiment is whether enough and good enough seed documents actually can be supplied by users in an operational situation.

The work done by Kwok represents an attempt to bypass the problem of having to specify the query in the form of a citation, while at the same time "…retaining the use of the citing/cited relationship between documents." (Kwok, 1985b, p. 166). In an ambitious series of papers (Kwok, 1975; 1984; 1985a; 1985b) it is proposed that the titles from cited and citing documents may be used to enhance document representations and ultimately improve retrieval performance. The idea can be seen as an implementation of Salton's initial idea (1962; 1963), although it is not put forward as such. It is presented in the context of automatic classification systems, and a theoretical justification is given for the use of cited titles in IR within the probabilistic IR theory

(Kwok, 1984; 1985a). In two small studies of the medical domain the effects on the statistical quality of clusters are examined (Kwok, 1975; 1985b). Generally the clusters are improved by inclusion of the titles of cited and citing documents. In spite of the explicit IR focus no attempt is made by Kwok to assess the effect on retrieval performance. After testing the idea on two small test collections (the CACM and CISI collections) using the SMART system Salton and Zhang (1986) found that many useful single words can be extracted from cited and citing titles, but also many of doubtful value. As they found no obvious way of separating them, Salton and Zhang conclude that the method is not reliable enough to be incorporated in operational IR systems. Despite these findings ISI adds keywords generated automatically from cited titles to their document representations. The so-called KeyWords Plus consist of single words and phrases up to three words long intended to function mainly as a recall enhancing device (Garfield, 1990; Garfield, 1993). KeyWords Plus are extracted from the cited titles by unpublished algorithms that select the "most significant" terms based on frequency information and various NLP procedures (Garfield, 1990, p. 297). In a small study Qin (2000) compares MeSH terms with KeyWords Plus for 400 records. However, no actual IR experiments on the effectiveness of KeyWords Plus have been reported so far, and we have no knowledge of the effect of these on retrieval performance. With the proliferation of scientific documents in full text, experiments on larger test collections could be valuable in assessing the value of cited titles as a document representation.

Another example of the use of citation data in automatic classification for IR is the experiments carried out by Shaw. Two effects of using term-based representations versus representations made from both references and citations were studied: the effects on the quality of the generated clusters (Shaw, 1990a; 1991a), and the effects on IR performance (Shaw, 1990b; 1991b). The cystic fibrosis (CF) test collection was used in all the experiments[22]. The term-based representations were derived from Medical Subject Headings (MeSH). Both of the citation-based representations were citation codes very similar to those in ISI's citation databases. The experiments consist of

[22] The CF test collection contains 1239 documents on cystic fibrosis from MEDLINE with codes added for references and citing documents and has 100 queries with extensive relevance judgements from several assessors on each topic (Shaw, Wood and Tibbo, 1991).

optimising a number of variables in relation to the relevance judgements: The database was partitioned into clusters for a range of variable values, and all clusters were evaluated as answer sets to all queries, but only the best obtainable results were reported. Overall the results showed that term and citation-based representations each could make statistically meaningful clusters on their own, and that the retrieval performance of the citation-based representations were comparable to, or marginally superior to what was obtainable with the term-based ones. An interesting feature of the experiments is that they introduced a weight based on the inverse document frequency (IDF) of the index keys. By setting a threshold on the weight the *exhaustivity* of the document representation could be controlled. The consequence of a low threshold was that only those index keys (both term and citation-based) which occurred relatively rarely in the database were included. In effect a degree of *selectivity* was introduced with respect to which references and citations (as well as MeSH terms) to include in the document representation. Better IR performance was obtained at low thresholds for both term and citation-based representations, but the citation-based representations yielded more consistent performance across thresholds, and superior performance at high thresholds. A separate study where the term and citation-based representations were combined resulted in better performance than with each constituent representation (Shaw, 1991b). On the whole, Shaw's experiments support the results obtained by Salton (Salton, 1971). It is important to note, however, that the approach cannot be implemented in an operational IR system because of the method chosen for the experiments. The reported results are for the best possible clusters only, but no method for selecting this in response to a query is considered in the experiments, which is a considerable weakness of Shaw's studies.

### *4.4.2 Experiments in operational settings*

A number of researchers from the two sub-traditions in the user-oriented tradition in IR have shown interest in the ways that end users may interact with citation data during their search for information. Ellis' (1989) study of the information seeking patterns of social scientists revealed that citation search strategies were used by all interviewed in the study: backward chaining was a major part of their seeking activities, and forward chaining, while not generally known, was used by a significant number of those who were aware of them. Based on these observations Ellis recommends that IR systems should be designed to support citation search strategies. A wide variety of investigations of the relative merits of different kinds of indexing and search strategies have been

carried out in the operational online R&D community. A few of these studies, such as those by McCain and Pao and associates reviewed below, focus on the contributions offered by citation searching in ISI's citation databases in comparison with subject searching in domain specific databases. The main goal has been to study the pragmatic use of citation data in end users' interaction with online systems for the fulfilment of their own information needs.

Based on practical experience Pao notes that subject searching from domain specific online databases and citation searching from ISI's citation indexes often result in different documents for the same request (Pao, 1984). She proposes that the differences between subject and citation searching should be compared and analysed in order to gain insights that may benefit online searchers. Analysing the number of relevant documents retrieved in response to nine requests in the medical behavioural sciences McCain (1989) compared the results of the two approaches to searching. In addition to three domain databases on DIALOG®, an average of 4.7 seed documents were searched using forward chaining in SCI and SSCI. Average recall values based on the union of the two sets of search results were calculated: On average, 57% of the relevant documents were identified in the domain databases, 33% in the citation databases, and 10% in the overlap between them. McCain notes that: "Citation retrieval was not equally successful for all topics, however, and there is a large variation in performance within the nine topics." (McCain, 1989, p. 113). From the examples given by McCain part of the reason for the variation in the success of the citation searches seems to be due to inadequate seed documents, e.g., because they were too recent to have received citations. Following her initial proposal, Pao conducted a minor study (Pao and Fu, 1985) before carrying out a larger two part investigation into the differences between the two types of searching. The first part was a pilot study using a small pharmaceutical in-house database with added references (Pao and Worthen, 1989). Real requests for information received from users of the database were used in the study, but the seed documents and relevance assessments were provided by experts. Overall, much better results were obtained by subject searching, both with respect to recall and precision. This is explained as being partly due to the limited size of the database, and partly due to a very thorough subject indexing. Pao and Worthen note that the identification of good seed documents is crucial: "A different "seed" could drastically alter the recall and precision measures of the citation searching…" (1989, p. 234). In several instances they had to request additional seed documents from the experts, and they note that there is no

apparent method to identify the "best" seed document in relation to a specific information need *a priori*. Similar to McCain (1989) only a small overlap was found: 4% on average. A separate analysis of the overlap showed that the overlap had a precision of 83% on average. This interesting finding was tested further in the second part of the experiment. This was a larger scale field study taking place in a realistic operational setting: Patrons of four different libraries were asked to evaluate the outcome of online searches, carried out in MEDLINE by intermediaries on their behalf, in return for a citation search on the topic in SCI free of charge (Pao, 1993). The patrons supplied one to three seed documents (with 90% of users only supplying one) for a total of 89 topics, and assessed the results for relevance on three grades. In contrast to McCain (1989) documents not indexed in both databases were eliminated from the analysis, resulting in a comparison on more equal terms. On average the MEDLINE searches yielded a precision of 56% and relative recall of 77%. For SCI the corresponding figures were 60% precision and 33% relative recall. 1466 extra documents were retrieved by SCI only, adding 24% more relevant documents to the total pool of relevant documents. There was an overlap of only 4.8% on average between all papers retrieved by the two approaches, with 28% of the topics not having any overlap. These overlaps had, however, an average precision of no less than 92%, confirming the results of the pilot study. Pao analysed the odds that the documents in the overlaps would be relevant as opposed to not relevant compared to either approach alone. She found that the documents in overlaps had 6.4 times higher odds of being partially relevant or relevant and 8.4 higher of being relevant. As discussed below the relatively small overlaps with high precision levels found in several of the reviewed studies are not surprising when seen with the theory of polyrepresentation in mind. In comparison to McCain (1989), Pao found a smaller overlap and a smaller number of relevant documents added by citation searching. Pao speculates that this may be partly due to McCain's success in obtaining a larger number of seed documents.

### *4.4.3 Other uses of references and citations in IR*

The value of references and citations for IR, partly documented by the experiments in the previous two sections, has made several designers include them in their systems.

The $I^3R$ prototype (*I*ntelligent *I*ntermediary for *I*nformation *R*etrieval) designed by Croft and Thompson exploited both references and citations as part of its many novel features

(1987). They were used in the graphical browsing expert when a retrieved document was presented to the user for judgement: the group of references in the document as well as the group of documents citing the retrieved document were displayed in a browse map as links, together with a link to a group of documents retrieved by term-based nearest neighbour clustering. In addition, the strength of the three types of links in the network was used to recommend new documents that the user might want to examine. The document representations including references and citations were generated from the CACM test collection (Fox, Nunn and Lee, 1988).

The increasing availability of scientific documents in full text has spawned a number of different initiatives to improve the access to these documents by exploiting the properties of full text. The increase in scientific documents in full text is a direct consequence of the fact that most scientific journals are now produced and typeset electronically. The result is that many journals are available in electronic form as well as print, and there is a growing number of journals in electronic format only. The change to an electronic production process at the major publishers has been facilitated by the Standard Generalized Mark-up Language (SGML) and the development of standard Document Type Definitions (DTD) like ISO 12083:1994 for facilitating authoring, interchange and archiving of publications like books and journals. This makes it possible to produce both print and electronic versions of a publication in the same process, independently of the final layout, by formatting the text in SGML mark-up. The text with mark-up can then be interpreted differently for different purposes, for instance to produce proofs for the print version, PDF and HTML files for an online version, and to extract data to generate indexes for an IR system.

One example is the Digital Library Initiative[23] (DLI) at the University of Illinois at Urbana-Champaign (UIUC), in which more than 40,000 articles from different publishers in the areas of engineering and physics have been made searchable in a single virtual collection (Schatz et al., 1996). The articles are formatted in SGML from which a variety of data types have been extracted directly for different retrieval and interface purposes, for instance to allow searches in author names, formulas, or figure captions etc. As a novel feature, the cited authors, titles and journal names are extracted directly

[23] See http://dli.grainger.uiuc.edu/idli/idli.htm

from the mark-up and may be searched separately. Apart from this use, the citation data are not exploited further in DLI. However, the implementation in DLI demonstrates that citation data may be extracted without great difficulty from documents with high quality mark-up when these are available. Thereby they could be exploited for more complex purposes in IR such as those reviewed in Section 4.4.1. However, the fact that the DLI is only accessible at UIUC because of copyright restrictions underlines the problems of getting access to large collections of scientific documents with high quality mark-up even for research purposes.

A different approach to the exploitation of citation data from full text documents in IR has been taken in the CiteSeer (Giles, Bollacker and Lawrence, 1998) and Rosetta (Bradshaw and Hammond, 1999) IR systems. Both of them crawl the WWW to identify freely available scientific publications as postscript or PDF files. These are then converted to text, parsed and added to the database of the systems. As these files are flat in the sense that no mark-up or formatting demarcate specific data types in the documents, specialised parsing techniques have to be applied to extract the desired information. In both CiteSeer and Rosetta particular attention is paid to references and citations in retrieval. The stated purpose of *CiteSeer*[24] is to create a citation database automatically from PS and PDF files (Giles, Bollacker and Lawrence, 1998). In CiteSeer papers are downloaded from the WWW, parsed to identify the references as well as the context where they are mentioned in the body text, and stored in a database. In addition, the full text is indexed and a range of browsing modes is supported. The papers may be searched by citation search strategies, or accessed by conventional term-based methods. CiteSeer is designed as an autonomous agent that retrieves papers in PS or PDF that are likely to be research papers from other web search engines (Bollacker, Lawrence and Giles, 1998). The agent uses heuristics to locate research papers and is invoked by a user who defines the area of interest by supplying broad keywords. Research papers are identified as those containing a bibliography. A public version of CiteSeer covering the computer science literature and with many features implemented is maintained at the NEC Research Institute[25]. A number of tests on small collections of references were run to determine the best way to identify citations to the same article as

[24] CiteSeer is also called ReserachIndex.

[25] See http://citeseer.com

the references routinely contain errors and appear in a number of very different formats in the downloaded papers (Lawrence, Giles and Bollacker, 1999). The best performing algorithm considered the whole reference as one string (without identifying specific subfields such as the cited author or title) and matched the citations after calculating a similarity measure between citations based on words and two term phrases. The phrases were defined as any two successive terms that occur in any section of the reference containing more than three terms (where a section was delimited by comma or full stop). The result was that the algorithm is sensitive to term ordering within the sections, but not between sections. No attempt to stem the terms using conventional methods was made. The algorithm first sorted all references by length, then for each reference it identified the group of references with the highest number of matching words. If the similarity measure between the group and the reference was above a threshold the reference was added to the group. If not, a new group was formed with the reference. Each of the resulting groups were then considered to refer to one document, and the longest reference in each group was used as representative of the group.

The three main uses of references and citations in CiteSeer are for backwards and forward chaining, as automatically generated hyperlinks, and to extract the citing context. Of these three cases only the last is not possible in ISI's citation indexes. When a document has been retrieved, its references and the documents citing it are linked automatically to the retrieved document if they are in the database. If the retrieved document has been cited, the context surrounding the citation in the full text is extracted and displayed in the interface as an indication of how the retrieved document has been received and discussed by later publications. The display consists of a window of a few sentences around the spot where the retrieved document is mentioned. Other uses of references and citations in CiteSeer include an automatically generated histogram showing the distribution of citing articles on years (excluding self citations), and an automatically generated list of related documents similar to that in Web of Science based on an implementation of bibliographical coupling. The recommendation of similar documents was based on a variant of the *tf*idf* weighting scheme applied to the citations: the "Common Citation × Inverse Document Frequency" (CC×IDF) was used to order the set of documents with at least one shared reference to the retrieved document currently under investigation. The IDF component "…assumes that if a very uncommon citation is shared by two documents, this should be weighted more highly than citations made by a large number of documents." (Bollacker, Lawrence and Giles,

1998, p. 5-6). No normalisation for document length is done on the CC component. Except for the small tests of the algorithm to identify citations to the same article, no formal tests of the performance of CiteSeer or the effect of CC×IDF weighting has been carried out so far. Goodrum et al. have examined the differences between samples drawn from CiteSeer and the computer science literature in SCI (2001). CiteSeer includes a large proportion of conference papers, also among the highly cited papers in the database. "Documents in both databases overwhelmingly cite books and book chapters, followed closely by journal articles." (Goodrum et al., 2001, p. 669). The major difference is that there is a higher proportion of citations to conference papers among the highly cited in CiteSeer (15%) vs. SCI (3%).

In the *Rosetta* system (Bradshaw and Hammond, 1999) the citation context in the full text of the citing article is used, not only in the interface as in CiteSeer, but also as a representation of the cited document. The idea can be seen as an extension of one proposed by Kwok (1984) and tested by Salton and Zhang (1986), and also an extension of the idea behind ISI's KeyWords Plus. Instead of the cited and citing titles, however, the potentially more precise and directly related citation context is used for document representation. Only preliminary tests of the idea have been carried out by the authors and it is not known it will affect IR performance (Bradshaw and Hammond, 1999; 2001; 2002).

## 4.5 Summary statements and discussion

As the number of search tools available to scientists continue to increase references and citations will probably maintain their position as an important part of scientists' practical information seeking behaviour. As shown in Section 4.4 and 4.4.3, references and citations have, from time to time, been considered for various purposes in IR research. A number of characteristics of their use for IR purposes can be learned from these studies:

1. Whatever the underlying motivations are for the selection of particular references by particular authors, it *is* possible to use references and citations to retrieve relevant documents, often with good results;
2. The predominant use of references and citations has been at the document level, i.e., as identifiers of the cited document;

3. By their nature references and citations are fundamentally different kinds of representations from the standard term-based ones. Representations generated from references and citations most often therefore retrieve other documents than term-based representations in response to the same information need, as well as relatively few overlapping documents;
4. When the performance of references and citations is compared with standard term-based representations in IR experiments, references and citations tend to increase precision at the expense of recall;
5. When both types of representations are applied, the combination often yields considerably better performance than the term-based representations alone;
6. Although retrieval based on references and citations can achieve good performance on average, large variation in performance across topics often occur, especially in operational studies;
7. The seed documents used to represent the information need in a forward chaining are crucial for the performance of references and citations in IR;
8. The increasing availability of scientific documents in full text provides new opportunities for exploiting references and citations in IR.

As shown by the analysis of references and citations as alternative representations in Section 4.3 there is no single explanation of why they are useful in IR (characteristic 1). Motivations for giving references have been analysed from normative and social constructivist positions, but neither of them succeeds in explaining the behaviour observed and the reception of citations by individual authors as registered in a citation index. It is difficult to get any further than to say that the act of giving references is part of the scientific discourse and is guided to a certain extent by its explicit and implicit norms and values. Together this mixture of factors make up the semantic relationship, which is utilised when searching for scientific documents by means of references and citations. This can presently be done by using a range of different search strategies, which for the main part involve some sort of forward chaining, a feature unique to citation indexing. The persistence of ISI's citation databases as well as the results of studies such as Ellis' (1989) are indications that references and citations continue to be useful for IR purposes.

As alternative representations, references and citations appear to be very specific indicators of documents content and subject matter. This is in line with Small's conception of references and citations as concept symbols (Small, 1978). In most

studies and applications, references and citations are represented by a citation code identifying a particular document (characteristic 2). An alternative that has not been explored to any great extent is to use parts of the citation code for retrieval purposes, e.g., the cited author or cited journal as unit instead of whole reference strings. Rare examples of this include Ding's use of static author and journal co-citation maps for online pre-search query formulation and expansion (2000), and Lin, White, and Buzydlowski's maps for online searching generated in real-time from author co-citation data (Lin, White and Buzydlowski, 2003).The empirical results behind characteristic 3 and 4 may partly be explained by the practice of representing references and citations as a whole document. This practice is in line with Small's theory of references as symbols for particular concepts in the course of an argument.

That a very high proportion of relevant documents was found in the overlap between term and citation-based representations (characteristic 5) is not surprising from a cognitive point of view. This can be seen as the result of producing an overlap between two very different types of representations: references and citations on one side, and controlled descriptors assigned by human indexers as well as words from titles and abstracts on the other side. The cognitive aspect derives from the fact that *several* different intellectual interpretations of the documents contribute to locating the documents in the overlap, i.e. the MEDLINE indexers' intellectual analysis resulting in descriptors, the author's own perceptions represented by title and abstract words as well as citations given by other authors. With such representations a small overlap with a high proportion of relevant documents would indeed be anticipated.

Variation in performance between queries in IR experiments is not unusual neither in laboratory experiments nor in operational experiments. Several of the reviewed studies, however, report significantly larger variation in performance with citation-based representations (Characteristic 6). Much of the variation seems to be due to the nature of the seed documents used. McCain (1989), Pao and Worthen (1989), and Pao (1993) all remark on the importance of good seed documents (characteristic 7), as does Salton (1971). A major difference between the experiments is that the laboratory experiments generally use a larger number of seed documents than the operational studies. The importance of acquiring good seed documents in a forward chaining becomes even more important when there are only few of them as in the operational studies: the well documented inconsistencies in human interpretation have a marked influence on the

results. If references are used as concept symbols (Small, 1978) this may explain these results: many symbols may be chosen to represent a concept, but if only a few seed documents are used in the forward chaining the outcome becomes very dependent of these. If more are used the chance of more consistent performance is higher. The most central issue of concern in the utilisation of references and citations in IR is therefore to identify methods that can improve the selection of seed documents. That is, to find ways in which to improve the seed documents that are used as representations of the user's information need. The increasing availability of scientific documents in full text might provide new opportunities to test the use of references and citations in IR, either with highly structured documents from publishers or with free available flat files acquired from the internet (Characteristic 8).

In all of the above, the basic unit that is being discussed and analysed in relation to references and citations has been the *document*, either as referring document or as cited document. Even in research evaluation the aggregations of citations counted, e.g., for departments or countries, are based on the citations to documents published. This may be regarded as a consequence of the particular citation code created by ISI. This code has determined many of the proposals and implementations reviewed above. Järvelin, Ingwersen and Niemi (2000) argue for improved citation indexes. They demonstrate an interface that can generate an enhanced citation index through advanced data modeling techniques directly from bibliographic references. With the availability of scientific documents electronically in full text new possibilities are offered, because more features of citations can be exploited based on the full text references in these. For instance, all cited authors would be available as well as all cited titles (of articles, journal, books, conference names etc.)

# 5 The boomerang effect

As demonstrated by the scattered studies reviewed in chapter 4, references and citations can be utilised as representations of documents and utilised in IR systems. However, references and citations are very rarely utilised directly as identifiers of document content in best match systems neither in IR research, nor in the operational IR systems designed to aid retrieval of scientific information. In spite of the promising results obtained both in laboratory and operational IR settings, CiteSeer remains one of the very few exceptions to this (Giles, Bollacker and Lawrence, 1998). A number of reasons may be given for this. In the system-driven IR research a major obstacle has probably been the problem of matching the citation-based document representation with the information need. In tests of ad hoc IR systems the information need has most often been expressed as one or more seed documents in some sort of citation code. Both the laboratory and the operational experiments showed that the selection of good seed documents is crucial, but in the laboratory setting no end users are available to supply seed documents. And although some of the early test collections consist of scientific documents and contain citation data that can be used for document representation, none of them explicitly include citation data as part of the requests. Except for Salton's (1971) neat but slightly unrealistic use of the base documents in the Cranfield test collection as seed documents, the system-driven IR research on references and citations has therefore been concentrated on the document representations alone, e.g., by adding cited titles to document representations (e.g., Kwok, 1985b), or by using the citation data for clustering purposes (e.g., Shaw, 1990b). Another factor that may have shifted the focus away from scientific documents in IR research is the start of the TREC experiments in the early 1990s (Harman, 1993). Here one of the intentions was to have a dramatic increase in the size of the test collection. This was achieved by using a large number of full text documents, but it meant that is was necessary to include such documents as were available electronically in full text. Therefore TREC includes news paper articles, federal reports, web pages etc., but no scientific documents in full text or with citation data incorporated so far, probably because these are much harder to get hold of free of charge. This situation may change as more and more scientific documents become available electronically in full text as discussed in the previous chapters. So far the recently established INEX initiative is the only IR initiative with a

corpus consisting of a fairly large amount of scientific articles in full text including references (Gövert and Kazai, 2003).

In this chapter a new method for exploiting citation data in IR is proposed, based on the theory of polyrepresentation (Ingwersen, 1996). The intention behind the so-called boomerang effect is to eliminate the need for the user to specify seed documents, and yet retain the potential advantages of references and citations as alternative representations in IR. With a natural language query as starting point, the boomerang effect *automatically* selects and weights seed documents for use in a forward chaining that can be submitted as an ad hoc query to a best match IR system. Different approaches to selection of seed documents, including an automatic one based on polyrepresentation, are considered in Section 5.1, followed by the proposal of the boomerang effect in Section 5.2. A small pre-experiment with a Boolean version of the boomerang effect is reported in Section 5.3. The experiences gained from the pre-experiment form part of the proposal of a best match boomerang effect Section 5.4, Section 5.5 contains summary statements and a discussion. The implementation of the best match boomerang effect is described in Chapter 6.

## 5.1 Identification of "good" seed documents

As demonstrated in Chapter 4 the availability of good seed documents is crucial in any retrieval strategy based on forward chaining. An important question is, therefore, what characterises a good set of seed documents for a given information need? Based on her experiences from the study of differences between descriptor and citation retrieval, McCain reasoned that effective citation retrieval requires: a) that there exists at least a few key papers on the topic of the information need, b) that the importance of these papers is generally recognised by researchers in the field, c) that the norms of scholarship in the field require citation of these key papers, and d) that sufficient time has passed since the publication of the key papers to generate a body of citing work. (McCain, 1989, p. 113). Except for explaining low performance for one of her information needs as a consequence of too recent seed documents, McCain does not investigate these prerequisites further, or propose methods to investigate or incorporate them in future experiments. Pao refrains from attempting to assess the appropriateness of the seed documents supplied by users in her studies (Pao and Worthen, 1989), and the prerequisites proposed by McCain (1989) do not seem to have been studied in any

dept since. At present we therefore have very limited knowledge of what characterises a good seed document.

One reason for the lack of such studies may be that investigators have seen no other options than to rely on the users' ability to supply appropriate seed documents. Indeed, in her field study Pao found that the patrons, as seekers of scientific information, could all supply at least one seed document, although very few provided more (Pao, 1993). This was also the case in the pilot study, but some of the seed documents returned very few documents from the forward chaining (Pao and Worthen, 1989). To this problem no other solution was found than requesting additional seed documents from the experts. Thus it may seem that even in operational settings, where searches are mediated and assisted by experienced information specialists, it can be hard to ensure that the given seed documents will be appropriate for retrieval purposes. As noted by Pao and Worthen there is a parallel between term-based and citation-based retrieval in that "…there is no absolute ideal search statement…" neither for term-based queries nor for citations-based ones (1989, p. 234). However, it remains an open question whether the interview techniques employed by information specialists, e.g., to overcome the label effect (Ingwersen, 1982) and extract a more detailed version of the information need from users, will be at all useful when obtaining seed documents from users. As demonstrated in Chapter 4 the research into the nature of giving references does not present any definite answers as to why authors select particular references, and our understanding of the phenomenon is quite limited. Furthermore, as there have been no studies of different methods of obtaining seed documents from users, our knowledge of what influences this choice can be described as almost nonexistent.

The number of seed documents used per information need is interesting as it could explain the considerable variation across different information needs experienced, e.g., in the operational studies reviewed in Section 4.4.2 above. If references function as concept symbols (Small, 1978) there might be many different ones of these that can match a particular information need. In combination with the fact that although references must function in a rhetorical context there are very large degrees of freedom for authors to chose particular references, this is a good explanation of the often low levels of recall found in earlier studies: if only a few seed documents are used to represent the information need they may succeed in achieving high precision, but it is not very likely that high recall will be achieved on average. Because an author can

select between many different references to represent the same idea, there is too little chance of retrieving a major part of the documents (and thus achieve high recall) that deal with the idea, when only a few seed documents are used.

Since good seed documents are crucial when utilising references and citations in IR it is desirable to explore principles for obtaining these. Such studies may take three forms:

a) Users may be studied in their contexts, and factors that influence the choice of seed documents can be investigated. This is in line with Cronin's conclusion on his investigation into the citation process (1984).
b) If a few seed documents are available these could be expanded from the network of references and citations, e.g., by selecting the nearest neighbours from a clustering of the citation data prepared in advance. Methods similar to those used by, e.g., Ding (2000), could be modified to achieve this. Salton's use of all the references from the base documents in the Cranfield collection as seed documents is also a simple example of this approach (1971).
c) If no seed documents are available an attempt may be made to identify them with the term-based query as starting point.

The proposed method, the boomerang effect, is described in the next section. The boomerang effect takes as a starting point that no seed documents are available (case c). Thereby the method is not dependent on the users' abilities to provide good ones, and can be used in the cases where the users are not familiar with the research field.

## 5.2 The proposed method: the boomerang effect

The proposed method is an attempt to automate the citation cycling strategy described in Section 4.2.1 by removing the need for the user to supply seed documents. Instead the information need expressed as a natural language request is automatically translated into citation codes by exploiting the network of references and citations. Thereby no seed documents need to be provided by users – only the natural language request as in traditional IR systems. The number of seed documents can also be increased considerably. A natural language request is probably a more realistic starting point than demanding that users provide seed documents. On the other hand the automatic transformation into seed documents introduces more uncertainty as to whether the seed documents identified will be good representations of the information need. In the

citation search strategies described in Section 4.2.1above this uncertainty is reduced by the user who intellectually selects a few promising seed documents, for instance among the references in documents retrieved using free text terms or descriptors in the uncontrolled and controlled subject search (type d and e). These seed documents then form the basis of a backward chaining or a cycling search.

The particular approach taken to identify seed documents in the boomerang effect is based on the theory of polyrepresentation. The intention is to reduce the uncertainty automatically by emphasising those citations that occur in the overlaps between documents identified by different cognitive and functional representations. As the boomerang effect is based on the theory of polyrepresentation, different versions of it with varying degrees of structure can be constructed along the polyrepresentation continuum (See Section 3.2.3 and Figure 3.3above). Two versions at each end of the continuum are presented below. Figure 5.1 illustrates an exact match version where distinct overlaps are identified. This exact match version was implemented and tested in the pre-experiment reported in Section 5.3. Section 5.4 presents a best match version of the boomerang effect. This is illustrated in Figure 5.6. Here the individual overlaps are not explicitly identified – instead the citations are weighted and submitted as a weighted query against a best match IR system. The best match version was tested in the main experiment, which is reported in Chapter 6 and 7.

### *5.2.1 An exact match boomerang effect*

The basic form of the boomerang effect is a cycling search from documents dealing with the information need, involving a backward chaining from documents identified in several different representations, and a forward chaining only from the overlaps of the sets generated by the backward chaining. This three-step cycling process and the identified overlaps are exemplified in Figure 5.1 below in the form of Venn diagrams. The example is based on an exact match IR system similar to those at the commercial online hosts where searching is done in specific indexes with Boolean operators and adjacency operators. Successful retrieval from such a system requires that the natural language expression of an information need is translated into the search language used by the system. Documents are in the form of bibliographical records similar to the ones in ISI's citation indexes, including the reference lists of the documents.

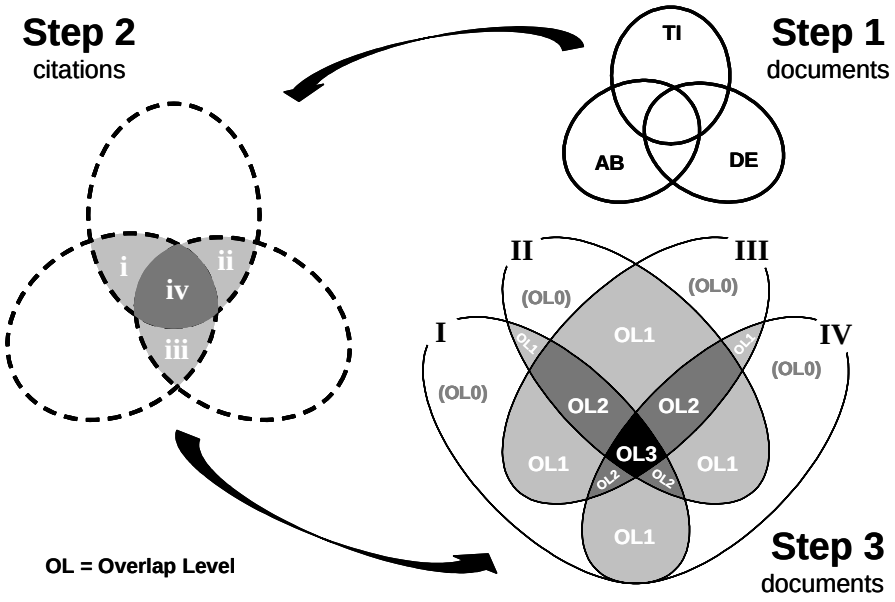



**Figure 5.1. Example of the exact match boomerang effect with 3 initial representations. In Step 3, four sets of documents (I-V) citing the citations contained in the overlaps in Step 2 (i-iv) are retrieved, using the citations as seed documents. The citations in Step 2 are extracted from documents retrieved by term-based queries in Step 1. (Modified from Larsen and Ingwersen, 2002).**

**Step 1** consists of identifying sets of *documents* on the topic of the information need from a number of different cognitive and functional representations of the same corpus of documents. In the example in Figure 5.1 three representations are searched using Boolean combinations of natural language terms: the title (TI), abstract (AB) and descriptor (DE) fields. Title and abstract both originate from the author, but they are functionally different. The descriptors are controlled index terms assigned by an indexer, and therefore of a different cognitive origin compared to title and abstract. Following the hypothesis in the theory of polyrepresentation, which is that the more different the representations generating the overlaps are cognitively, the higher the probability that the overlap contains a large proportion of relevant documents. Different search tactics and formulations may be used for each representation depending on the type of representation. For instance, adjacency operators are likely to be applied when searching in titles and abstracts, whereas a thesaurus might be consulted when deciding which descriptors to use. The result of Step 1 is a result set of documents from each representation on the topic on the request. If polyrepresentation alone were to be tested, the relevance of documents occurring in and outside the different overlaps would be analysed paying particular attention to overlaps generated by cognitively different

representations. In the boomerang effect the overlaps in Step 1 are not explicitly examined, but they influence Step 2 and 3. Note that at least two sets are needed in Step 1 to generate any overlaps subsequently. If available, more may be used, which will generate a rather complex set of overlaps at different levels, but for the sake of simplicity only three are showed in the example in Figure 5.1.

For **Step 2** the *references* from each of the results sets from Step 1 are extracted into a pool for each set. This is similar to a backward chaining, but on all the references contained in the documents. However, the documents pointed at by the references are not retrieved (this is indicated by a dashed line around the pools in Figure 5.1) – only the citation codes representing the documents are processed. Instead the pools of references in Step 2 are used to select seed documents for a forward chaining. In the exact match boomerang effect the distinct overlaps between the pools are identified by locating those references that occur in more than one pool. Inspired by the theory of polyrepresentation the idea is that although there may be many inherent inconsistencies in the reference lists (because of various motives for selecting references, as discussed in Chapter 4) much of this uncertainty will be reduced if only the overlaps between pools are used as seed documents. By the extraction into pools and the identification of overlaps, the references contained in the documents in Step 1 are transformed into *citations*. Therefore the pools and overlaps in Step 2 may be regarded as a temporary citation index created at search time. In Figure 5.1 the three pools in Step 2 generate four overlaps – three in which the references are present in two of the pools (overlaps i, ii and iii – marked in light grey), and one in which they are present in all three pools (overlap iv – marked in dark grey). The identification of overlaps is such that a citation cannot occur in more than one overlap, and there are hence no duplicate citations between, for instance, overlap iii and iv. Although overlaps between documents are not analysed explicitly in Step 1 the references from any document in an overlap in Step 1 will automatically be included as citations in an overlap in Step 2. The overlaps in Step 1 can be used to weight the citations in Step 2, for instance to select between the citations.

In **Step 3** the citations in the overlaps from Step 2 are used as seeds in a forward chaining. Hereby *documents* are retrieved that refer to the seed documents in overlap i, ii, iii and iv (Figure 5.1). The four sets of retrieved documents (I, II, III and IV) may be regarded as belonging to one, large set as in, e.g., McCain (1989) or Pao (1993).

Alternatively they may be kept separate and split into new sets of overlaps depending on which of the four overlaps from Step 2 they refer to as shown in Figure 5.1. The result of the exact match boomerang effect is 1) the documents in Step 3 that refer to the seed documents in the overlaps in Step 2; and 2) the overlap structures in which the documents in Step 3 can be ordered. In Figure 5.1 the overlaps in Step 3 are marked in different shades of grey, and the *overlap level* (OL) of each overlap is indicated. Combining several document sets can generate overlaps at different levels: An overlap on OL1 is generated by two sets, whereas it takes three sets to generate an overlap on OL2. The documents in set I-IV that are not part of any overlap are referred to as being on overlap level 0 (OL0). With this notation the maximum number of overlap levels is equal to the number of representations in Step 1. With three representations in Step 1 generating a maximum of four overlaps in Step 2 there is no less than a maximum of 11 different possible distinct overlaps in Step 3, with one OL3, four OL2, six OL1 as well as four sets that do not form part of any overlap (marked with OL0 in the figure)[26]. However, the maximum number of overlaps will rarely be reached in actual practice as shown by the pre-experiment (see Section 5.3). The overlaps may be thought of as forming a pyramid as shown in Figure 5.2: The sets identified by the forward chaining from Step 2 are found at OL0, with the overlaps generated by these sets as levels on top. The number of documents in each overlap level will be equal to or less than that in the level below as a consequence of the Boolean operations used (except in some cases for OL0 where the number of documents may be less than that in OL1 because the overlaps are subtracted from the entire set of retrieved documents, see, e.g., Table 5.6).

A wide range of topics would be covered both central and peripheral to the information need if *all* the documents represented by the citations in Step 2 were considered. Following the hypothesis in the theory of polyrepresentation (see Section 3.2 above) the citations in the overlaps between the pools will probably point to documents that are more central to the information need because they are referred to by two, or more, different representations. These citations are therefore selected as seed documents for

[26] Drawing Venn diagrams (Venn, 1880) with $m$ concentric circles only is not possible for $m > 3$, and curves with other shapes have to be used (see, e.g., Ruskey (2001) for examples). Recently it has been proved that it is possible to draw Venn diagrams for any prime number of $m$ (see Cipra, 2003). In Figure 5.1 the ellipse solution for $m = 4$ in step 3 is from Rousseau (1998).

the forward chaining in Step 3 and the hypothesis in the best match boomerang effect is that a larger proportion of relevant documents will be found at higher overlap levels (see Figure 5.1 and Figure 5.2). Hereby the exact match boomerang in effect ranks the documents retrieved in Step 3 (which includes most of those retrieved in Step 1 – see below) in a number of strata constituted by the overlap levels.

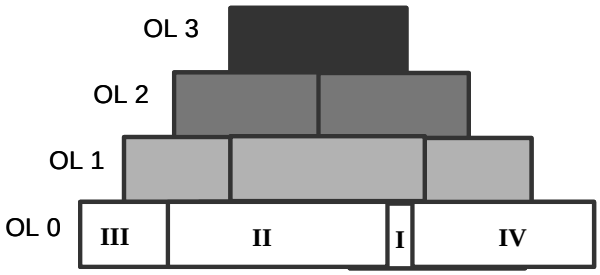



**Figure 5.2. Visualisation of the overlap levels in Step 3 of the exact match boomerang effect. The expectation is that the proportion of relevant documents will be greater at higher overlap levels. From Larsen (2002).**

This is similar to the idea behind the Boolean polyrepresentation search strategy in the appendix of Ingwersen (1996) discussed in Section 3.2.3 above, and may be said to simulate the ranked output from a best match IR system, except for the fact that documents are not ranked within each overlap level. Note that the only expression of the information need required for the boomerang effect is in natural language. No seed documents need to be specified in advance, and the seed documents used in the subsequent forward chaining are not selected intellectually. In the pre-experiment the processes to execute the best match boomerang effect were carried out manually using a range of different programs. As the processes are fairly simple and entirely logical without the involvement of human interpretation, a small programme could, given an experienced programmer, probably automate the operations necessary to execute the best match boomerang effect. Not shown in Figure 5.1 is the fact that any document found in the initial Step 1 will be included in Step 3 provided that the document includes at least one reference in one of the overlaps in Step 2. This follows immediately from Proposition 1 in Egghe and Rousseau (1990, p. 230ff ), which is based on Kochen's work on citation networks (Kochen, 1974). In Egghe and Rousseau (2002) this is generalised into a basic cycling theorem. The documents initially retrieved in Step 1 could be removed, and only the documents unique to Step 3 inspected. It should, however, be far more rewarding to examine those of the documents that are

*retrieved by both approaches*, especially those documents that are placed in the overlaps in Step 3. The documents in Step 3 not retrieved by Step 1 could display a different set of characteristics and point to documents that are on different subjects, but using the same theories, methods, apparatus etc. on different subject areas. Indeed, because the match in Step 3 is done entirely on citation codes the retrieved documents may be using quite different terminology than expressed in the information need and still be relevant. Exploring these documents more closely could unveil interesting links to other subject areas and fields hitherto not noticed. Indeed, a consequence of the translation of the information need into citation codes is that the boomerang effect functions as a *query expansion tool* as extra documents not otherwise identified from the natural language expression of the information need will be retrieved. Unlike traditional query expansion, where for instance terms from relevant documents or synonyms to search keys are added to the query (see, e.g., Rocchio, 1971; or Kekäläinen, 1999), the additional documents retrieved by the boomerang effect are identified via the network of references and citations. Thus the end result of the exact match boomerang effect is the additional documents retrieved, but also the overlap levels that the documents from both Step 1 and Step 3 can be ordered in.

The boomerang effect is named so because it may be said to resemble the path taken by a boomerang[27]. Although not dependent on seed documents, the boomerang effect is of course dependent on the quality of the initial searches (stated as the 'quality-in-quality-out principle' by Croft and Thompson (1987). As noted by Pao and Worthen (1989) so are any search strategies based on natural language terms. As only citation codes are processed and matched in Step 2, measures need to be taken to minimise negative effects of the errors and inconsistencies in the citation codes that were discussed in Chapter 4. In the cystic fibrosis test collection the references and citations of the 1.239 documents included in the collection were corrected for errors and standardised

[27] The following description of the use of a boomerang (the returning type) was kindly provided by an aboriginal present at the ISSI2001 conference: "A skilled, and fortunate, boomerang hunter will, after having approached his prey from the windward side, throw the boomerang, hit the animal, and bring it down. For a skilled and not quite so lucky boomerang hunter the boomerang will return after missing the prey (for instance if the prey suddenly ducked for some reason), spinning three time above the head of the hunter before dropping gently to the ground at his feet."

intellectually by the investigators (Shaw, Wood and Tibbo, 1991). Such quality control is only practically viable on small document collections – for larger ones some form of automatic procedure needs to be considered. A potential problem in the exact match boomerang effect is that retrieved documents without references in Step 1 will not influence the documents retrieved in Step 3 even if they are relevant, nor can any document without references be retrieved in Step 3. However, documents without references do not tend to be a problem when using a citation index in practice (Garfield, 1979). In the citation index generated from the corpus used in the main experiment there were 16 references per document on average for those documents containing references, and 40% of the documents in the corpus entirely without references (See Section 6.2.4 below). Because the boomerang effect is based on sets of documents and pools of citations the problem will in practice very rarely prevent the boomerang effect from retrieving a large number of documents in Step 3. In the pre-experiment, for example, 88 documents were retrieved on average per information need in Step 1, and 431 documents were returned per information need on average in Step 3 in the overlaps. It is therefore appropriate to consider different tactics that can either make the result set larger or smaller. These include:

- Restricting the documents in Step 1 and 3 to those that are published recently.
- Only considering documents in Step 2 of a certain age, e.g., restricting the citations to the oldest or youngest half respectively[28].
- Considering only higher overlap levels, either in Step 2, Step 3 or both.
- Reducing the references in Step 2 to those that are cited above a certain threshold, either globally in the citation index or by the documents in Step 1.
- Increasing or decreasing the number of representations searched in Step 1.

The last tactic depends on the number of representations available in the corpus. Although in theory there is no upper limit on the number of representations that can be used in Step 1, the maximum number of overlaps in Step 3 increases rapidly with additional representations. With $n$ representations in Step 1 the maximum number of overlaps in Step 3 is given by:

[28] Sumner (1995) examined the effect of such a partition on the clustering quality of a subset of the cystic fibrosis test collection. His results showed that were no discernable effects on clustering quality of partitioning the citations by age in a small collection.

$$O(n)_{\max\ Step\,3} = 2^{\left(2^n-(n+1)\right)} - (n+2) \qquad (1)$$

As Formula 1 is a double exponential function there will be maximum of 2,042 overlaps with 4 representations and no less than 67,108,857 overlaps with 5 representations. Although a lot of the overlaps both in Step 2 and 3 will be empty in practice it will be prohibitive to identify all overlaps with more than 4 representations, even if the exact match boomerang effect was automated by a program. The partition into overlap levels illustrated in Figure 5.2 simplifies the structure of Step 3, but preserves the partial ranking of the retrieved documents to a certain extent. If all sets generate overlaps in Step 2 and 3 the number of overlaps levels will not exceed the number of representations used in Step 1. With the reduction of Step 3 to overlap levels, specific overlaps become indistinguishable as does the connection between an overlap and the representations in Step 1 that gave rise to it. This is not incompatible with the intention of the theory of polyrepresentation. Even if the distinct overlaps were kept the double distillation of overlapping citations in the exact match boomerang effect has the effect that any overlap in Step 3 will be the result of overlaps between all representations in Step 1, albeit in a slightly different mix. If the distinct overlaps are reduced to overlap levels as proposed it is consequently not feasible to study the effect of a single representation on performance by examining specific overlaps[29]. An alternative way to study such effects is to exclude the representations one by one and test the overall performance of the remaining possible combinations.

## 5.3 Pre-experiment

A small empirical pre-experiment was conducted in order to gain insights into the feasibility of implementing the boomerang effect in a larger experiment and to get indications of whether the hypothesised effects on performance will occur in a practical

[29] Differences between overlaps at the same OL might be expected though: All overlaps in set IV in step 3 (Figure 5.1) can be seen as more important because they refer to citations contained in the super overlap in step 2. By the reduction into overlap levels these differences are also lost.

case[30]. As no best match IR facilities that could handle the necessary operations were available at the time, an exact match version of the boomerang effect near the structured pole of the polyrepresentation continuum was implemented, i.e. the pre-experiment was carried out using methods similar to the ones in the user-oriented studies of, e.g., McCain (1989) and Pao (1993).

Based on the principle of polyrepresentation the main hypothesis in the pre-experiment is that the exact match boomerang effect has the potential of enhancing the performance of IR systems by identifying relevant documents. The following research questions were investigated in the pre-experiment:

1. Is it possible to retrieve relevant documents via the networks of references and citations without specifying seed documents in advance, using an implementation of the exact match boomerang effect?

This question investigates the basic mechanics of the exact match boomerang effect to discover if it can at all retrieve documents using the proposed approach. If, for instance, no or very few overlapping citations were found in Step 2 the exact match boomerang effect might fail to retrieve any documents in Step 3.

2. Is there a larger proportion of relevant documents in the overlaps in Step 3, especially in those at higher overlap levels?

This question concerns the main hypothesis in the boomerang effect, and is a test of whether the theory of polyrepresentation, as implemented in the boomerang effect, can be shown to have a positive effect on IR performance. If this is the case the documents retrieved in Step 1 and those additional documents retrieved in Step 3 may be combined to improve IR performance.

### *5.3.1 Methods and data*

Three work tasks with accompanying search statements have been constructed in cooperation with a domain expert who acted as test person. The test person was an experienced researcher at a hospital and the work tasks reflect his current research interests into osteoporosis (brittleness of the bones). Following an unstructured interview about his current research interests and typical information seeking activities

[30] The pre-experiment has previously been described in Larsen and Ingwersen (2001) and Larsen (2002).

three work tasks were constructed. An initial exploration of the PubMed database[31] was carried out together with the test person in order to obtain words and phrases that could be useful as search keys. The test person is a frequent and experienced user of PubMed, and formulates complex Boolean queries easily, although he did not explicitly take advantage of restricting search keys to certain fields, e.g., the title or MeSH fields. A favourite search strategy is the 'building block search strategy' (Harter, 1986) submitted to the "All fields" of PubMed, which includes words from titles, abstract and assigned MeSH terms. The work tasks consist of two parts: a short, verbal description of the situation leading to the information and details of the information need itself, as well as a Boolean search statement constructed in cooperation with the test person. The Boolean search statements consist of the terms and blocks originally proposed by the test person as well as extra synonyms. The extra synonyms were included after initial tests in order to facilitate matches in all the representations to be used in Step 1[32]. Figure 5.3 shows an example of a work task. All three work tasks are given in Appendix 1.

**Work task 3:**

As a researcher your main research interest is osteoporosis (brittleness of the bones). Earlier research has shown that osteoporosis is influenced by a variety of factors, such as physical activity, age and sex of patients, whether women are pre- or postmenopausal etc. You are interested in finding evidence from the literature that can indicate how genetic factors affect osteoporosis. You give special attention to studies that compare monozygotic twins with biovular twins.

(BMD OR bone mass OR BMC OR osteoporo* OR fracture* OR BUA OR SOS OR QUS) AND (genetic OR heredit* OR polymor* OR mutat*) AND (twin* OR gemel*)

**Figure 5.3. Sample work task as used in the pre-experiment. Consists of a verbal formulation of the work task and the actual search statement used.**

No single available IR system was capable of handling the operations needed in the pre-experiment. Therefore several available systems and programs were combined in a

[31] PubMed is the web-based version of Medline with more than 12 million bibliographic records. PubMed can be accessed free of charge at http://www.ncbi.nlm.nih.gov/entrez/

[32] An analysis of the Boolean search statements proposed by the test person revealed that based on experience he often ensured high precision in his searches by the combination of words that might appear such as title words or MeSH terms with very specific terms or abbreviations that almost only occur in the abstract.

simulation of the boomerang effect as it might operate if implemented in a single program. The initial searches in Step 1 were carried out in SCI through Web of Science. SCI was chosen because it covers the medical sciences and includes the references of the documents, and access through Web of Science was chosen because records could be downloaded free of charge for further manipulation. The initial retrieval of documents in Step 1 does not display polyrepresentative properties. These are invoked in the following steps. Queries were constructed using Boolean operators and submitted to the TOPIC field, which contains terms from titles, abstracts, author keywords and KeyWords Plus (this is similar to what may be found in the basic index in online databases, e.g., at Dialog). Except for the titles, the fields cannot be searched separately in Web of Science. Therefore it was chosen to run the queries including the extra synonyms against the TOPIC field, download all resulting hits from Web of Science, and then identify the documents for each representation in Step 1 separately offline. In retrospect this strategy was not entirely successful as a considerable part of the documents could only be identified in the basic index, and not by any of the individual fields (see below). This echoes the difficulties experienced by Madsen and Pedersen (Madsen and Pedersen, 2003). As discussed above in Section 3.2.3a possible solution to this it is to consider further expansion and adapt this to each representation.

As an operational, multidisciplinary database, Web of Science contains a very large number of records. In order to reduce the resulting sets to a manageable size for the pre-experiment, only documents published in 1999 and 2000 were included in Step 1 and Step 3. The set of returned documents for each work task was downloaded from Web of Science. As most of the documents retrieved in Step 1 will be retrieved in Step 3 (see Section 5.2.1 above), these documents were analysed separately in order to study differences between these and the additional documents identified by the exact match boomerang effect. The following six different aggregations of the documents retrieved in the pre-experiment will be analysed:

Step 1 The documents retrieved in Step 1, distributed on representations (264 documents in total for all three work tasks excluding overlaps between representations)

Step 1' The subset of documents from Step 1 sampled for relevance assessments (146 documents)

Step 1" The subset of documents from Step 1' that were also identified in Step 3 (135 documents)

Step 3 All the documents identified in Step 3 (4774 documents)

Step 3' The subset of documents from Step 3 that were assessed for relevance. Consists of Step1" and Step 3" (135+143 = 278 documents)

Step 3" The subset of documents from Step 3 that were sampled for relevance assessments, excluding documents from Step 1" (143 documents)

The major part of the analysis below will focus on the samples in Step 1" and Step 3".

The identification of overlaps between citations in Step 1 was handled by a series of programs: First, the documents were downloaded and imported into the database software *Reference Manager*[33]. The following representations were considered for use in Step 1: Titles (TI), abstracts (AB), author keywords (DE) and Keywords Plus (ID). The first three are functionally different representations generated by the author. The author keywords were discarded because only 64% of the records contained these. Although Keywords Plus are automatically generated from the citation indexes by ISI they were included as they are not generated by the author, and represent the closest thing to a cognitively different representation that could be extracted from the Web of Science records. Documents that could be retrieved by each of the three representations (TI, AB and ID) were identified in Reference Manager using the Boolean search statements from the work tasks. A significant part of the documents (46%) could only be retrieved in the basic index consisting of index keys from both TI, AB and ID. Rather than discarding these documents from the pre-experiment they were considered to be an extra representation and added as a fourth representation in Step 1 (called CO

[33] Reference Manager 9 from ISI Researchsoft was used (http://www.refman.com).

for the combination of other fields). The rationale behind this is that although the documents in the CO representation do not display properties that include them in any of the other representations separately they nonetheless matched the Boolean query in the basic index. There is thus a reasonable chance that they may be relevant. The documents in the CO representation may be said to represent the kind of documents that might be retrieved in Step 1 if the full text could be searched.

The references were extracted from the downloaded records into citation pools and pre-processed by *Bibexcel*[34], because there are many inconsistencies in the very short and dense cited reference (CR) string in ISI's citation databases. Figure 5.4 shows an example of the inconsistencies that can occur. The CR strings for three articles by the renowned Danish biochemist Jens F. Rehfeld are shown. The two articles in *Journal of Biological Chemistry* (Rehfeld, 1978a; 1978b) are a two-part study which appeared in the same issue (volume 253, issue 11 which begins page 4016 and 4022 respectively). By verifying the bibliographical data a number of formal errors can found in the 28 CR strings identified for the two articles, several of them occurring in one string: 13 in the volume numbers, 3 in the page numbers, 2 in the cited year, as well as 4 errors in the initials of the cited author. In addition, 5 different cited work (CW) forms have been used, one page number does not cite the first page, and 3 strings do not contain all data, making it impossible to verify them. The third article (Rehfeld, 1981), which appeared in *American Journal of Physiology* (volume 240, issue 4, pages G255-G266), displays similar errors with a new type added because of alphanumeric characters in the page numbers in the original article. This uncommon feature seems to have confused the referring authors or the ISI indexers, and in four cases the first page number is included as part of the cited work. Note that when considering the citation frequencies (column 2 in Figure 5.4) the majority of the citations are represented by a few CR strings especially for highly cited articles.

Inconsistencies in the CR string have the potential of being detrimental the boomerang effect both in Step 2 and Step 3. Therefore an attempt was made to standardise the string in Bibexcel to facilitate a high quality matching. As many variants occur in the

[34] Bibexcel is a tool for offline bibliometric analyses constructed by Professor Olle Persson. It may be downloaded free of charge at http://www.umu.se/inforsk/

cited work and the cited page parts of the CR strings they were reduced to cited author, cited year and cited volume, and all citations without these three elements removed (10 % of the citations). This represents a compromise between reducing the inconsistencies in the cited reference strings and ensuring that not too many irrelevant false matches will be made. Figure 5.5 shows the strings in Figure 5.4 after processing in Bibexcel. It may be observed that the number of strings has been reduced, but that not all variants have been collated. The main effect of the standardisation is that all citations to articles by the same (first) author from the same year and appearing in the same volume are collated, if these data have been cited correctly. A possible source of errors is that citations to different articles published in the same journal in the same year will be represented by the same citation code. However, this will probably only happen very rarely, and even when it does there is a good chance that the subject matter of the articles is similar in nature as in the example in Figure 5.4. As can be seen in Figure 5.4 errors occur in almost all the elements and there are no simple solutions to rectify this. Higher quality matches might be obtained by a combination of rule-based tactics, such as those used in artificial intelligence, and approximate string matching techniques.

| List # | freq. | CR string |
|---|---|---|
| E1 | 1 | CR=REHFELD J, J BIOLOGICAL CHEM |
| E2 | 1 | CR=REHFELD J, 1978, V10, P4022, J BIOL CHEM |
| E3 | 1 | CR=REHFELD J, 1978, V25, P4016, J BIOL CHEM |
| E4 | 2 | CR=REHFELD JA, 1978, V235, P4022, J BIOL CHEM |
| E5 | 1 | CR=REHFELD JD, 1978, V253, P2016, J BIOL CHEM |
| E6 | 2 | CR=REHFELD JF, J BIOL CHEM |
| E7 | 3 | CR=REHFELD JF, J BIOLOGICAL CHEM |
| E8 | 1 | CR=REHFELD JF, 1978, V153, P4016, J BIOLOGICAL CH |
| E9 | 1 | CR=REHFELD JF, 1978, V252, P4016, J BIL CHEM |
| E10 | 1 | CR=REHFELD JF, 1978, V252, P4016, J BIOL CHEM |
| E11 | 2 | CR=REHFELD JF, 1978, V252, P4022, J BIOL CHEM |
| E12 | 1 | CR=REHFELD JF, 1978, V253, P4002, J BIOL CHEM |
| E13 | 291 | CR=REHFELD JF, 1978, V253, P4016, J BIOL CHEM |
| E14 | 46 | CR=REHFELD JF, 1978, V253, P4016, J BIOLOGICAL CH |
| E15 | 2 | CR=REHFELD JF, 1978, V253, P402, J BIOL CHEM |
| E16 | 2 | CR=REHFELD JF, 1978, V253, P4020, J BIOL CHEM |
| E17 | 630 | CR=REHFELD JF, 1978, V253, P4022, J BIOL CHEM |
| E18 | 80 | CR=REHFELD JF, 1978, V253, P4022, J BIOLOGICAL CH |
| E19 | 1 | CR=REHFELD JF, 1978, V253, P4023, BIOL CHEM |
| E20 | 3 | CR=REHFELD JF, 1978, V253, P422, J BIOL CHEM |
| E21 | 1 | CR=REHFELD JF, 1978, V256, P4016, BIOL CHEM |
| E22 | 1 | CR=REHFELD JF, 1978, V256, P4016, J BIOL CHEM |
| E23 | 1 | CR=REHFELD JF, 1978, V263, P4022, J BIOL CHEM |
| E24 | 1 | CR=REHFELD JF, 1978, V283, P4022, J BIOL CHEM |
| E25 | 1 | CR=REHFELD JF, 1978, V283, P4022, J BIOLOGICAL CH |
| E26 | 1 | CR=REHFELD JF, 1982, V253, P4022, J BIOL CHEM |
| E27 | 1 | CR=REHFELD JH, 1978, V252, P4016, J BIOL CHEM |
| E28 | 1 | CR=REHFELD JR, 1976, V253, P4022, J BIOL CHEM |
| ... | | |
| E32 | 2 | CR=REHFELD JF, 1981, AM J PHYSIOL |
| E33 | 1 | CR=REHFELD JF, 1981, P240, AM J PHYSIOL |
| E34 | 1 | CR=REHFELD JF, 1981, P255, AM PHYSL SOC G |
| E35 | 1 | CR=REHFELD JF, 1981, V240, E255 AM J PHYSIOL |
| E36 | 1 | CR=REHFELD JF, 1981, V240, G225 AM J PHYSIOL |
| E37 | 59 | CR=REHFELD JF, 1981, V240, G255 AM J PHYSIOL |
| E38 | 51 | CR=REHFELD JF, 1981, V240, G255, AM J PHYSIOL |
| E39 | 8 | CR=REHFELD JF, 1981, V240, G255, AM J PHYSL |
| E40 | 1 | CR=REHFELD JF, 1981, V240, G266 AM J PHYSIOL |
| E41 | 7 | CR=REHFELD JF, 1981, V240, PG255, AM J PHYSIOL |
| E42 | 4 | CR=REHFELD JF, 1981, V240, P255, AM J PHYSIOL |
| E43 | 3 | CR=REHFELD JF, 1981, V240, P255, AM J PHYSL G |
| E44 | 12 | CR=REHFELD JF, 1981, V240, P6255, AM J PHYSIOL |
| E45 | 1 | CR=REHFELD JF, 1981, V240, P755, AM J PHYSIOL |
| E46 | 2 | CR=REHFELD JF, 1981, V240, P9255, AM J PHYSIOL |
| E47 | 1 | CR=REHFELD JF, 1981, V249, G255, AM J PHYSIOL |
| E48 | 1 | CR=REHFELD JF, 1982, V240, G255, AM J PHYSIOL |
| E49 | 1 | CR=REHFELD JF, 1984, V240, G255, AM J PHYSIOL |
| E50 | 1 | CR=REHFELD JH, 1981, V240, P255, AM J PHYSIOL |

**Figure 5.4. Example of inconsistencies in cited reference (CR) strings. Each CR string is preceded by its list number and the number of times it is cited (Source: SCI, ISI, 2003)**

```
CR=REHFELD J, 1978, V10?
CR=REHFELD J, 1978, V25?
CR=REHFELD JA, 1978, V235?
CR=REHFELD JD, 1978, V253?
CR=REHFELD JF, 1978, V153?
CR=REHFELD JF, 1978, V252?
CR=REHFELD JF, 1978, V253?
CR=REHFELD JF, 1978, V256?
CR=REHFELD JF, 1978, V263?
CR=REHFELD JF, 1978, V283?
CR=REHFELD JF, 1982, V253?
CR=REHFELD JH, 1978, V252?
CR=REHFELD JR, 1976, V253?
...
CR=REHFELD JF, 1981, V240?
CR=REHFELD JF, 1981, V249?
CR=REHFELD JF, 1982, V240?
CR=REHFELD JF, 1984, V240?
CR=REHFELD JH, 1981, V240?
```

**Figure 5.5. Example of cited reference strings after standardisation in Bibexcel.**

The matching of citations across pools to identify the overlaps in Step 2 was done by a query in *Microsoft Access*, and all overlaps in Step 2 were kept separate and labelled in Reference Manager, in order to be able to identify each of the sets and overlaps in Step 3 afterwards. This resulted in a fairly large number of citations in Step 2 (240 citations on average per work task), even though only documents published in 1999 and 2000 were included in Step 1 (see Table 5.2). As the amount of data was too large to handle in the simulation it was decided to reduce the size of the pools. Thus only the most frequently cited seed documents in each pool in Step 2 were used for the forward chaining. This selection of seed documents is very similar to the calculation of weights proposed for the best match boomerang effect (See Section 5.4 below). The remaining seed documents were submitted to SCI. As a large amount of seed documents were to be submitted and the overlaps between them identified, the web-based interface in Web of Science could not be used. Instead the seed documents were submitted to the online version of SCI at Dialog via *DialogLink* in batches for each pool. This was followed by an online identification of the documents in overlap levels. As can be seen from Table 5.3 below, quite a large number of documents were retrieved in Step 3: 88 documents were retrieved on average per information need in Step 1, and 431 additional documents were returned on average in the overlaps in Step 3, in spite of the fact that both Step 1 and 3 were restricted to documents published in 1999 and 2000. A sample of these was

therefore drawn and downloaded for further analysis as described below. The time used to execute the simulation of the exact match boomerang effect was approximately one day of full time work per information need after these had been collected from the test person. Much of this time was spent handling the overlaps and ensuring their consistency and correctness. While this simulation involves quite a large amount of manual work the processes themselves are simple and straightforward, and could easily be automated.

The procedure described above was used to retrieve documents for each of the work tasks. The retrieved documents, in the form of printed records from SCI including abstracts, were presented to the test person who judged them for relevance in relation to the information need behind the work tasks. Based on the test person's own information behaviour it was agreed that up to 100 documents would be assessed for each work task. Approximately half of the assessed documents were drawn from Step 1 and the other half from those documents retrieved in Step 3, excluding documents already retrieved in Step 1'. The sampled documents were mixed randomly within each work task so that the test person could not identify their origin. Table 5.1 shows the number of documents retrieved from Web of Science for each work task, and their distribution on representations in Step 1. Because of overlaps between the representations the sum of the four representations is larger than the total. It can be seen that a significant part (38%) of the documents was retrieved by the combination of representations only (CO) and that work task 1 and 2 did not retrieve any documents by Keywords Plus (ID). This reduces the number of possible overlaps in Step 2 and 3. The references of all 264 documents were used in Step 2, but a maximum of 50 documents in Step 1 from each work task was presented to the test person for relevance assessments. These were drawn randomly from work task 1 and 2. All 49 documents from work task 3 were presented to the test person for relevance assessment.

**Table 5.1. Number of documents initially retrieved for each work task (WT) without duplicates between representations (Total), and their distribution on representations in Step 1. Referring documents published in 1999-2000 (Source: Web of Science, ISI, 2001)**

| | TI | AB | ID | CO | Total |
|---|---|---|---|---|---|
| WT 1 | 10 | 72 | 0 | 41 | 119 |
| WT 2 | 3 | 17 | 0 | 77 | 96 |
| WT 3 | 4 | 14 | 15 | 21 | 49 |
| Total | 17 | 103 | 15 | 139 | 264 |

Table 5.2 shows the distribution of the number of unique citations in the pools after processing in Bibexcel. Table 5.2 also shows the cumulated number of citations in the overlaps for each work task (in total and using the concept of overlap levels on Step 2) as well as the actual number of seed documents used for the forward chaining, selected by weighting the most cited from each pool in every overlap. The most frequent citation in a pool in Step 2 was mentioned 51 times in that pool. All combinations of the pools did not generate overlaps in Step 2: In work task 1 and 2 there were three overlaps (out of four possible), in work task 3 there were six overlaps (out of eleven possible).

**Table 5.2. Distribution of unique citations in each of the representations in Step 2 on work tasks (WT), the distribution and total number of unique citations in overlaps, as well as the number of references used in the forward chaining. (Source: Web of Science, ISI, 2001)**

| | TI | AB | ID | CO | "OL1" | "OL2" | "OL3" | Total (OL1-3) | Used |
|---|---|---|---|---|---|---|---|---|---|
| WT 1 | 29 | 1,048 | - | 708 | 280 | 23 | - | 303 | 98 |
| WT 2 | 105 | 572 | - | 2,259 | 224 | 15 | - | 239 | 96 |
| WT 3 | 50 | 439 | 254 | 386 | 94 | 67 | 17 | 178 | 66 |

Table 5.3 displays the number of documents retrieved as a result of the forward chaining in Step 3. It may be seen that work task 1 and 2 cannot achieve overlap levels larger than 2, because they generate only three overlaps in Step 2. As expected, because of the Boolean operations involved, fewer documents were retrieved in the higher overlap levels. The only exception is work task 3 where more than half of the documents are placed in OL1. It may be noted that in practise there are no documents at OL5 in this study although it is possible because of the six overlaps at Step 3 in work task 3. The last row in Table 5.3 shows the number of documents from work task 1 to 3 that were assessed by the test person distributed on overlap levels (Step 3"). A convenience sampling was used to ensure that a larger proportion of documents were selected from higher-level overlaps, and to keep the number of documents to be assessed by the test person within the agreed amount. Although the main interest is on documents contained in the overlaps in Step 3, a few documents were also sampled from OL0 in Step 3 in order to get an indication of the relevance of the documents outside the overlaps in Step 3. A total of 143 documents were sampled and mixed randomly with the 149 documents randomly drawn from Step 1 within each work task (Step 3"). A total of 289 relevance assessments were obtained for the three work tasks (3 documents were overlooked by the test person and excluded from the rest of the

analysis). The relevance of the documents was assessed in relation to the information need behind the work tasks using a scale with three degrees of relevance:

a) Definitely relevant
b) Maybe relevant
c) Not relevant

The scale is similar to those used by, e.g., Lancaster (1969), Saracevic (1969), Saracevic and Su (1989) as well as Pao (Pao and Worthen, 1989; Pao, 1993).

**Table 5.3. Distribution of documents on overlap levels and work tasks (WT) identified as a result of the forward chaining in Step 3, and the number of documents assessed for relevance (excluding documents assessed for relevance in Step 1'). Citing documents published in 1999-2000.**

| | OL0 | OL1 | OL2 | OL3 | OL4 | OL5 | OL1-5 | Total |
|---|---|---|---|---|---|---|---|---|
| WT 1 | 1,322 | 152 | 3 | - | - | - | 155 | 1,477 |
| WT 2 | 1,701 | 203 | 66 | - | - | - | 269 | 1,970 |
| WT 3 | 457 | 688 | 136 | 44 | 2 | 0 | 870 | 1,327 |
| Sum | 3,480 | 1,043 | 205 | 44 | 2 | 0 | 1,294 | 4,774 |
| Assessed (Step 3") | 41 | 68 | 23 | 10 | 1 | 0 | 102 | 143 |

### *5.3.2 Analysis of results*

The main measure of performance in the pre-experiment is *precision*. Precision is defined as the number of relevant documents retrieved, divided by the total number of documents retrieved (Cleverdon, Mills and Keen, 1966). The other commonly applied performance measure in IR evaluation, *recall,* has not been employed. Recall is defined as the number of relevant documents retrieved, divided by the number of relevant documents in the collection (Cleverdon, Mills and Keen, 1966). As the study is based on an operational IR system with several millions of documents, it has not been possible to obtain relevance assessments for all the documents in the collection, and recall has not been assessed. Relative recall such as calculated in Pao's studies (Pao and Worthen, 1989; Pao, 1993) where the union of output from two different strategies is used as a recall base has not been calculated as (nearly) all documents from Step 1 are included in Step 3. Precision takes values from 0 to 1 with high values as the ideal. In order to calculate recall and precision as defined above the relevance assessments must be binary. A method to calculate *generalized* recall and precision values on non-binary relevance data has recently been proposed (Kekäläinen and Järvelin, 2002b). This

method is used in the main experiment. However, since the data set in the pre-experiment is rather small this method will not be used here. Instead the data for all degrees of relevance will be reported, and analysed separately when needed. Based on the assumption that a user might want to examine documents assessed both as *Maybe relevant* and *Definitely relevant* it was chosen to merge these two categories into the category *All relevant* in the analyses. Precision in the pre-experiment is calculated as the number of documents in the merged category divided by the total number of documents assessed. No attempt was made to test the statistical significance of the results because only three work tasks from the same test person on the same general topics were used, and because the convenience sampling used to identify the documents in Step 3" does not lend itself to statistical testing.

A total of 160 out of the 289 assessed documents were judged to be relevant. This results in an overall precision of 55% across all works tasks, including documents assessed from Step 1' and Step 3". As can be seen from Table 5.4 and Table 5.5 there was, however, large variation among the work tasks and also between Step 1' and Step 3". Differences among the work tasks can be expected, as the nature of each of them is different, e.g., in their scope or specificity. There is a clear connection between the number of documents retrieved in Step 1 (Table 5.1) and the share of relevant documents in Step 1' (Table 5.4), but this connection does not extend to Step 3" (Table 5.5). The most obvious differences between Table 5.4 and Table 5.5 are the ones between the documents identified in Step 1' and those added by Step 3". The precision of the initial searches in Step 1' is very good, ranging from 48% to 91% with an average of 71.2% (Table 5.4). As an absolute value this is quite high, which is probably due to the fact that the work tasks derived from the test person are on very specific topics, and that the search statements reflect this. It is fair to say that the test person has optimised his search strategy to achieve high precision. The documents added by Step 3" (excluding the ones already retrieved in Step 1') do not display such a high precision, ranging from 28% to 63% with an average of 39.2% (Table 5.5). In both Step 1' and 3" there is a high precision in work task 2, whereas the added documents in work task 3 have a precision that is less than a third of those identified in Step 1' (28% versus 91%). It may also be noted that there is a marked difference in the degrees of relevance between the two steps: The share of Definitely relevant documents out of All relevant documents was 58% in Step 1' and 39% in Step 3".

**Table 5.4. Distribution of relevant documents on work tasks (WT) in Step 1'. Documents from overlap level 0 are included.**

| | Not relevant | Maybe relevant | Definitely relevant | All relevant | Total | Precision |
|---|---|---|---|---|---|---|
| WT 1 | 26 | 9 | 15 | 24 | 50 | 48.0% |
| WT 2 | 12 | 12 | 26 | 38 | 50 | 76.0% |
| WT 3 | 4 | 23 | 19 | 42 | 46 | 91.3% |
| All | 42 | 44 | 60 | 104 | 146 | 71.2% |

**Table 5.5. Distribution of relevant documents on work tasks (WT) in Step 3" (excluding documents assessed for relevance in Step 1'). Documents from overlap level 0 are included.**

| | Not relevant | Maybe relevant | Definitely relevant | All relevant | Total | Precision |
|---|---|---|---|---|---|---|
| WT 1 | 36 | 9 | 5 | 14 | 50 | 28.0% |
| WT 2 | 17 | 15 | 14 | 29 | 46 | 63.0% |
| WT 3 | 34 | 10 | 3 | 13 | 47 | 27.7% |
| All | 87 | 34 | 22 | 56 | 143 | 39.2% |

If the sample of documents retrieved in the initial search (Step 1') is used as baseline for the performance of the proposed citation search strategy, the latter does not appear to perform very well. However, this overall calculation does not take into account the overlap structure with its different levels and the expectation that the level of precision will rise together with the overlap levels. The results for all 3 work tasks distributed on overlap levels can be seen in Table 5.6 (Step 1") and Table 5.7 (Step 3"). The first five rows in the two tables show the overlap levels separately, and it can be observed that as the overlap level increases so does precision at each level increases in both cases when All relevant are considered. This also holds true for Step 1" when the documents assessed as either Definitely relevant or Maybe relevant are considered. For one level in Step 3", however, this is not the case: Only 2 Definitely relevant out of 10 documents were identified on OL3 (20%) whereas there were 6 out of 23 on OL2 (26%). Please note that only 135 of the 146 assessed documents from Step 1' were retrieved in Step 3 (Table 5.4 and Table 5.6). This is mainly a consequence of using the most cited references only in Step 2 as described above – had all references been used, all documents from Step 3 would have been retrieved, except those that do not have any

citations in the overlaps in Step 2. The study is primarily focussed on the documents contained in the overlaps, and this seems justified by the fact that the additional documents identified by Step 3" display a markedly lower precision at OL0 (17%) compared to the documents in the overlaps (OL1+2+3+4 = 48%). In Step 1" the difference is smaller: The documents on OL0 have a precision of 67%, whereas the documents in overlaps display a precision of 80.8% (OL1+2+3+4). This may be seen as a cautious confirmation that the documents retrieved by both strategies (i.e. are situated in the overlap between them) are more likely to be relevant, but can also be interpreted as an artefact of the very precise and focussed work tasks, which narrows the range of possible variation.

**Table 5.6. The distribution of relevant documents on overlap levels (OL), separately and cumulated in Step 1".**

| Overlap level | Not relevant | Maybe relevant | Definitely relevant | All relevant | Total | Precision |
|---|---|---|---|---|---|---|
| OL4 | 0 | 0 | 1 | 1 | 1 | 100.0% |
| OL3 | 0 | 2 | 3 | 5 | 5 | 100.0% |
| OL2 | 2 | 11 | 15 | 26 | 28 | 92.9% |
| OL1 | 16 | 15 | 28 | 43 | 59 | 72.9% |
| OL0 | 14 | 10 | 18 | 28 | 42 | 66.7% |
| OL 4 | 0 | 0 | 1 | 1 | 1 | 100.0% |
| OL 4+3 | 0 | 2 | 4 | 6 | 6 | 100.0% |
| OL 4+3+2 | 2 | 13 | 19 | 32 | 34 | 94.1% |
| OL 4+3+2+1 | 18 | 28 | 47 | 75 | 93 | 80.6% |
| OL 4+3+2+1+0 | 32 | 38 | 65 | 103 | 135 | 76.3% |

**Table 5.7. The distribution of relevant documents on overlap levels (OL), separately and cumulated in Step 3" (excluding documents assessed for relevance in Step 1').**

| Overlap level | Not relevant | Maybe relevant | Definitely relevant | All relevant | Total | Precision |
|---|---|---|---|---|---|---|
| OL4 | 0 | 0 | 1 | 1 | 1 | 100.0% |
| OL3 | 2 | 6 | 2 | 8 | 10 | 80.0% |
| OL2 | 10 | 7 | 6 | 13 | 23 | 56.5% |
| OL1 | 41 | 16 | 11 | 27 | 68 | 39.7% |
| OL0 | 34 | 5 | 2 | 7 | 41 | 17.1% |
| OL 4 | 0 | 0 | 1 | 1 | 1 | 100.0% |
| OL 4+3 | 2 | 6 | 3 | 9 | 11 | 81.8% |
| OL 4+3+2 | 12 | 13 | 9 | 22 | 34 | 64.7% |
| OL 4+3+2+1 | 53 | 29 | 20 | 49 | 102 | 48.0% |
| OL 4+3+2+1+0 | 87 | 34 | 22 | 56 | 143 | 39.2% |

Apart from the fact that the analysis of the sampled documents retrieved by Step 1 and those added by Step 3 has shown that there is a larger proportion of relevant documents in Step 1 in this study, it may also be observed that the proportion of Maybe relevant documents is larger than Definitely relevant among the documents added by Step 3 in almost every instance (Table 5.4, Table 5.5, Table 5.6 and Table 5.7). Thus the added, relevant documents do not seem to be obviously relevant at first glance, but nevertheless relevant enough to be examined further. As stated above, both the additional documents retrieved, as well as the overlap structure of all retrieved documents, should be seen as the result of the proposed strategy. The combined results (Step 3') can be seen in Table 5.8. From the viewpoint of a functioning IR system a number of possible tactics for the order in which to present the retrieved documents can be devised based on the results. The documents in higher-level overlaps could be presented to the user with the highest level first. For example, displaying the documents at overlap levels 2, 3 and 4 will result in a precision of 79%, with 34 documents presented from the initial subject search, and 34 of the additional. Moving further down to lower-level overlaps will of course decrease precision – especially when adding the at times large number of documents from lower-level overlaps (Table 5.3). The point is, however, that this descent can be made gradually.

**Table 5.8. The distribution of relevant documents on overlap levels (OL), separately and cumulated in Step 3'.**

| Overlap level | Not relevant | Maybe relevant | Definitely relevant | All relevant | Total | Precision |
|---|---|---|---|---|---|---|
| OL4 | 0 | 0 | 2 | 2 | 2 | 100.0% |
| OL3 | 2 | 8 | 5 | 13 | 15 | 86.7% |
| OL2 | 12 | 18 | 21 | 39 | 51 | 76.5% |
| OL1 | 57 | 31 | 39 | 70 | 127 | 55.1% |
| OL0 | 48 | 15 | 20 | 35 | 83 | 42.2% |
| OL 4 | 0 | 0 | 2 | 2 | 2 | 100.0% |
| OL 4+3 | 2 | 8 | 7 | 15 | 17 | 88.2% |
| OL 4+3+2 | 14 | 26 | 28 | 54 | 68 | 79.4% |
| OL 4+3+2+1 | 71 | 57 | 67 | 124 | 195 | 63.6% |
| OL 4+3+2+1+0 | 119 | 72 | 87 | 159 | 278 | 57.2% |

### *5.3.3 Discussion*

The pre-experiment was user-oriented in the sense that the information needs, the Boolean search strategies, as well as relevance assessments were provided by a domain expert. Two research questions were investigated in the pre-experiment.

The first research question concerns whether it is at all possible to retrieve documents without specifying seed documents in advance with the exact match boomerang effect. The result showed that this is indeed possible, in spite of several restrictions set or experienced in the execution: only documents published in 1999 and 2000 were included in Step 1 and 3, not all pools in Step 2 and 3 generated overlaps, and a threshold was applied in Step 2 to reduce the pools to a manageable size. Additional documents were retrieved in the overlaps in Step 3 for all three work tasks in the pre-experiment: on average 431 documents were added to the 88 found in Step 1 per work task (Table 5.3). This was achieved automatically without the intellectual selection of good seed documents necessary in previously described citation search strategies. Due to the procedures used, 93% of the documents from Step 1' were also retrieved in Step 3, and ordered together with all documents from Step 3 in an overlap structure.

The second research question investigates whether a larger proportion of relevant documents are found in the overlaps in Step 3' especially at the higher levels of the overlap structure as hypothesised in the exact match boomerang effect. The results show that of the sample selected for assessment documents *outside* the overlaps in Step 3" display noticeably lower precision (17%) compared to the documents in the overlaps (OL1+2+3+4 = 48%). The corresponding figures for Step 1" are 67% and 81%. Furthermore the precision at each overlap level is as a rule greater than or equal to the precision in the overlap level below in both Step 1" and Step 3". This also holds when each degree of relevance is analysed separately, except for a single case in Step 3" where OL2 has a greater share of Definitely relevant documents than OL3. Based on these results it is therefore recommended that the documents are displayed to the user in order of their presence in higher-level overlaps, so as to maximise the chances that as many relevant documents as possible will be presented first to a user. This takes full advantage of the exact match boomerang effect to increase the likelihood that precision is high in the first documents displayed and makes it possible to increase recall gradually in a controlled manner.

In contrast to a traditional Boolean system the result set may be expanded with the additional documents retrieved in Step 3. In the pre-experiment the sample of these of documents has a lower precision on average, but still a larger proportion of relevant documents in the higher level overlaps. When comparing the degrees of relevance in the sample there is a tendency towards a larger share of Definitely relevant documents in Step 1 than in Step 3. In spite of this the important point is that a much larger number of documents were retrieved in total in the overlaps in Step 3 (431 per query) compared to the initial queries in Step 1 (88 per query). Even though these additional documents might on average have a lower precision and fewer Definitely relevant documents, a given query can still be expanded in a controlled manner by returning the top overlaps first. For instance if the same pattern is found in all documents retrieved in Step 3 as in the sample the 431 additional documents found per query in OL1+2+3+4 will have a precision of 48%, and 84 additional documents found per query on OL1+2+3 will have a precision of 65%. Unfortunately, due to the inappropriateness of the sample for statistical testing we do not know the probability of finding the same patterns in the sample and in the whole population.

A number of factors need to be kept in mind when interpreting the results:

1. The analysis of the documents retrieved in Step 1 in response to the information needs showed that the information needs and Booleans queries were highly focussed and displayed a high precision on average.
2. No statistical testing was carried out on the document sets assessed for relevance. The results are therefore only valid for the assessed samples.
3. The nature of the work tasks in the study may influence the results in several ways. The initial subject searches were very focussed and displayed a very high level of precision. Thus as a result the extracted citations may be of a very high quality and therefore very well suited as input for the boomerang effect. This may bias the result in favour of the citation strategy, but it could on the other hand raise the baseline of the comparison so high that the exact match boomerang effect has little chance of matching the subject search, or showing if the theory of polyrepresentation can effectively reduce the inconsistencies and uncertainties as assumed.
4. The representations used in Step 1 were not ideal from a cognitive point of view, as they were only functionally different, and because only one representation could be extracted from the Web of Science records, which was not generated by the author. The KeyWords Plus are generated automatically from cited titles and may not be very strong indicators of document content compared to, e.g., intellectually assigned descriptors. If Step 1 had been carried out online, one could identify parallel sets of documents both with references (SCI) as well as intellectually assigned descriptors (MEDLINE) with the *Reverse Duplicate Removal* technique proposed by Ingwersen and Christensen (1997). Introducing such representations will probably result in smaller overlaps, but these might be of higher quality. When other types of representations are introduced there might be an increased need to adapt the queries to each representation, as experienced by Madsen and Pedersen (2003). In the pre-experiment it was attempted to expand the highly specific Boolean queries with synonyms so that documents could be found with each representation. This was not entirely successful as a considerable number of the documents could only be identified with the combination of representations (CO). This representation may have had a large influence on the results, for instance by generating large pools in Step 2 leading to larger overlaps.
5. Although mainly based on exact match an element of best match is introduced in Step 2 by the calculation of weights to make the threshold. It is not known how

this affects the results of the boomerang effect compared to using, e.g., all or a random sample of the citations in the overlaps.

In conclusion, the pre-experiment was successful in that it demonstrated the ability of the boomerang effect to retrieve relevant documents through the network of references and citations, and because it did identify larger proportions of relevant documents in the overlaps generated by the boomerang effect, especially at higher overlaps levels. The pre-experiment was closest to the structured pole of the polyrepresentation continuum, and made use of Boolean principles throughout except for in the selection of seed documents. The Boolean approach has two disadvantages in relation to the boomerang effect: *First*, with many representations (more than three) the number of overlaps and the effort required to handle them increases dramatically. *Second*, it is hard to compare the performance of the semi-ranked result of the boomerang effect to the ranked output from the best match systems that predominantly are being investigated in IR research. The purpose of the dissertation is to investigate references and citations in the context of these. Therefore, the exact match boomerang effect is abandoned in favour of a best match boomerang effect in the main experiment. This is described below.

## 5.4 A best match boomerang effect

The aim of the best match boomerang effect is to take advantage of frequency information as well as the theory of polyrepresentation to produce a ranked list of documents in Step 3. It has the same basic cycling form as the exact match version, but best match ranking principles are involved at each of the three Steps, placing the best match boomerang effect at the un-structured end of the polyrepresentation continuum. Figure 5.6 illustrates the best match boomerang effect with the same initial representations as in Figure 5.1. The documents in Step 1 and Step 3 are retrieved using a best match IR system. The cycling process is completed by a weighting of citations in Step 2 from which seed documents are selected as input to Step 3. In the best match boomerang effect the information need does not have to be formulated as a Boolean search statement, although it might be if the best match IR system used supports Boolean operators. Similar to the exact match version the documents must include the reference lists.

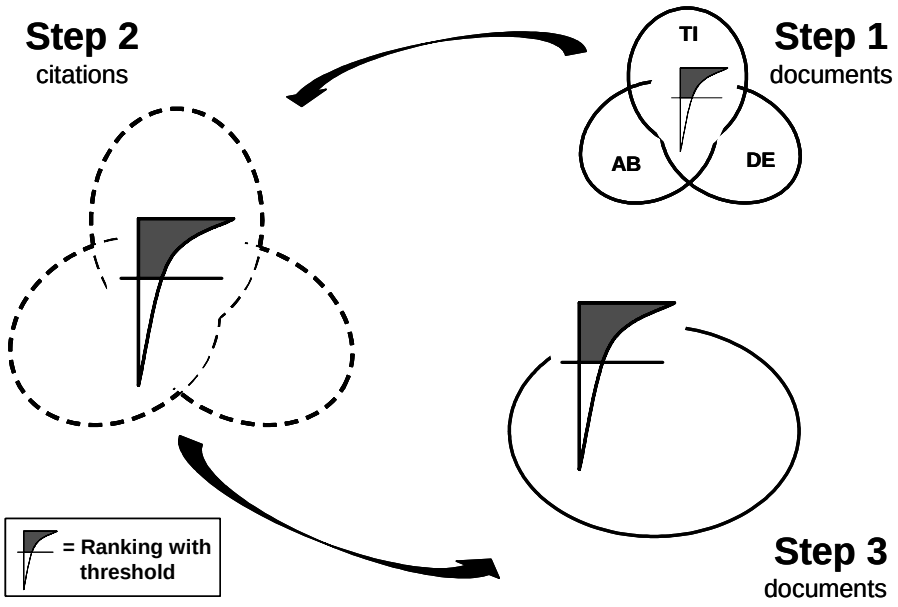



**Figure 5.6. Example of the best match boomerang effect. All steps involve some sort of ranking with a threshold cut-off. The citations in Step 2 are extracted from the top(n) documents returned by a best match run in Step 1. The citations are selected and weighted as illustrated by the matrix in Figure 5.7, and the top(n) percentile is submitted as a weighted query resulting in a ranked list of documents.**

**Step 1** consists of identifying sets of documents on the topics of the information need from different cognitive and functional representations of the same corpus. However, since the output of a best match system *per se* is a ranked list of documents containing all or just some of the search keys in the request, it is necessary to apply a *threshold* to the sets retrieved in Step 1. This is done to ensure the quality of the sets as it might otherwise be very low without the firm restrictions determined by the Boolean search formulations in the exact match boomerang effect. The result of Step 1 is a set of documents from each representation ranked algorithmically according to their expected relevance to the information need by the best match system.

In **Step 2** the references are extracted into pools of citations for each set in Step 1. Instead of only identifying the distinct overlaps as in the exact match boomerang effect, the frequency with which a given citation occurs in and across pools is exploited to weight the citations in Step 2. The matrix in Figure 5.7 is an example of how this weighting can be achieved. Normalised weights for a given citation *i* is calculated in

each pool *p,* and the weights added across pools into a final weight for the citation. The raw citation frequencies need to be normalised in order to maintain a balance between cognitively different pools: if there are great differences in the number of citations between two pools the overlap between them will be dominated by the largest pool if the calculation of weights is based on the raw frequencies of the citations in each pool. Thereby the idea of exploiting different cognitive structures is weakened because pools with a significantly smaller number of citations compared to the other pools will have very little influence on the final weights, and any effect of polyrepresentation is minimised. It is attempted to ensure a more balanced mixture between the pools in the best match boomerang effect by normalising the raw frequencies. This may be done in a number of ways. One might for instance consider how rare the citation is in the whole corpus and give higher weights to those citations that are rare in the corpus similar to the use made of the *idf* weight in CiteSeer (see Giles, Bollacker and Lawrence, 1998 or Section 4.4.3 above). The normalisation chosen for the best match boomerang effect is more simple (See Figure 5.7 below): The raw frequencies of citations are normalised for the *pool size* by dividing the frequency of occurrence of each citation *i* in a given pool *p* with the total number of occurrences of citations in that pool ($\sum fr_i$). This results in a *pool weight* for each citation *i* in each of the pools it occurs in. These pool weights are similar to the *tf* weights normalised for document length used in best match systems (See e.g., Section 6.2.1 below). The advantage of this normalisation is that a citation will be given a large pool weight if it occurs frequently relative to the size of the pool regardless if the pool is large or small. Similar to the exact match boomerang effect, only references occurring in the overlaps in Step 2 are considered in order to reduce the uncertainty and inconsistency related to different citer motivations. Therefore, citations which occur in one pool only are excluded. Consequently each of the remaining citations will occur in at least two pools and have several different pool weights. In accordance with the theory of polyrepresentation the pool weights for each citation *i* are *added* into one final weight.

The left-hand side of the matrix in Figure 5.7 maps the raw frequency of occurrence of five references ($i_{1\text{-}5}$) in three pools ($p_{1\text{-}3}$), and the right-hand side shows the pool weights as well as the final weights. Note that no weight is assigned to $i_5$ as it occurs in one pool only. Compared with the exact match boomerang effect, every citation in the overlaps in Step 2 of the best match boomerang effect is assigned a weight. Hereby the strict, but inflexible overlap structure in the exact match boomerang effect has been broken down

and replaced by a ranked list of citations with associated weights. In this way the top of the rank can consist of, e.g., citations with medium weights from all or most pools, or alternatively citations with large weights from a few pools.

| | *Frequency of occurrence of citations in the pools* | | | | | *Pool size* | *Pool weights (Citations weighted by pool size)* | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | $i_1$ | $i_2$ | $i_3$ | $i_4$ | $i_5$ | $\sum_{i=1}^{5} fr_i$ | $i_1$ | $i_2$ | $i_3$ | $i_4$ | $i_5$ |
| $p_1$ | 2 | 1 | 1 | 0 | 0 | 4 | $2/4$ | $1/4$ | $1/4$ | 0 | – |
| $p_2$ | 1 | 2 | 0 | 1 | 0 | 4 | $1/4$ | $2/4$ | 0 | $1/4$ | – |
| $p_3$ | 2 | 1 | 2 | 1 | 1 | 7 | $2/7$ | $1/7$ | $2/7$ | $1/7$ | – |
| | *Final weights for* $i_{1-5}$: | | | | | | **1.04** | **0.89** | **0.54** | **0.39** | – |

**Figure 5.7. Example of the calculation of weights for citations in the overlaps at Step 2, based on the occurrence of citations ($i_{1\text{-}5}$) in the pools ($p_{1\text{-}3}$). Modified from Larsen and Ingwersen (2002).**

In **Step 3** the weighted citations from Step 2 are submitted as seed documents to a best match IR system against an index containing the references of the documents. The weights calculated in Step 2 may be utilised for two purposes. Only the top-ranked citations may be selected as seed documents by applying a threshold to exclude citations with small weights in Step 2. The citations with associated weights can also be used as seeds and submitted as weighted search keys in an IR system that allows this, e.g., one based on the vector space model or the inference network model. A combination of both is possible as well. The end result of the best match boomerang effect is a list of documents with at least one of these seed documents, ordered in a single continuous rank by how many and how heavily weighted seed documents they contain.

The best match boomerang effect has a number of advantages over the exact match boomerang effect. The exponentially increasing number of overlaps that had to be handled in the exact match boomerang effect is not a problem in the best match boomerang effect, and it is thus easier to exploit many more representations. The best match boomerang effect produces a ranked output. The performance of the best match boomerang effect can therefore be compared directly to that of other best match IR techniques using standard procedures. A potential disadvantage is that the Boolean

control over the process is lost, which may lead to decreasing performance as experienced by Madsen and Pedersen (Madsen and Pedersen, 2003). When implemented at the un-structured end of the polyrepresentation continuum, the best match boomerang effect relies mainly on the best match system used in Step1 and Step 3 to provide the quality needed for the input.

## 5.5 Summary statements

The Chapter has initially identified the selection of good seed documents as the main challenge in the utilisation of references and citations in IR. Inspired by the theory of polyrepresentation (Ingwersen, 1996) the boomerang effect is presented as method that eliminated the need for the user to specify seed documents as a representation of their information need. Instead, the boomerang effect translates a natural language request automatically into seed documents that may be used in a forward chaining. The pre-experiment with a Boolean version of the boomerang effect confirmed that relevant documents could be retrieved with it, and that a larger proportion of relevant documents were found in the overlaps. Finally, a best match boomerang effect is proposed. This is tested in the main experiment as detailed in the following chapters.

# 6 Test data and methods

The boomerang effect may be tested in many ways. The document corpus has to consist of scientific documents in electronic form including references as noted in Chapter 5. The most suited format of the documents is a structured format such as SGML or XML, because the extraction of representations is greatly facilitated. Other alternatives might by considered, e.g., to use PDF files. However, specialised parsers are necessary to extract representations from such files, in particular the references. Such parsers have been shown to work satisfactorily in the CiteSeer project (Giles, Bollacker and Lawrence, 1998), but were not available for the dissertation work. The XML corpus in INEX was therefore used. Even with the well-developed parsing tools that are available for XML documents it was a major effort to extract representations for testing the boomerang effect. One of the consequences of this is that there was not sufficient time to develop a prototype interface so that users could be involved in the experiments. The evaluation of the boomerang effect therefore relies on the relevance assessments in the INEX test collection, and is system-driven is this respect as discussed in Chapter 2.

The purpose of this chapter is to describe the test setting in which the main experiment took place, the data used and the methods employed. Section 6.1 describes the Initiative for the Evaluation of XML retrieval (INEX) and the test collection created by the initiative as this is used in the main experiment. This is followed by an account of the used IR system and the test databases created, including the citation index constructed for the experiment in Section 6.2. Section 6.3 describes the test runs and baselines set up for the main experiment including the queries that were used. Section 6.4 discusses performance evaluation in IR experiments and describes the evaluation measures used, followed by Section 6.5 on statistical testing. The chapter concludes with summary statements in Section 6.6.

## 6.1 The INEX initiative

The Initiative for the Evaluation of XML retrieval (INEX) was announced in March 2002 with the purpose of creating a test collection for the evaluation of the effectiveness of XML retrieval. Interest in exploiting the structure of documents more directly in IR

has grown with the emergence and use of recognised standards for structuring documents, such as the SGML and XML standards (See, e.g., Chiaramella, 2001). While synthetic XML data can be used for tests of applications dealing with highly structured XML data similar to those in relational database management systems, applications for IR purposes have to be tested on real world documents (Fuhr et al., 2002). Prior to INEX no such test collections have been developed that are specifically designed for IR purposes, except for one with a small corpus of Shakespeare plays (Kazai, Lalmas and Reid, 2003). Therefore the overall objective of INEX is to create a test collection that will make it possible "…to assess a system's retrieval effectiveness, where effectiveness is measured as a system's ability to satisfy both content and structural aspects of a user's information need and retrieve the most specific relevant document components, which are exhaustive to the topic of request and match its structural constraints." (Gövert and Kazai, 2003, p. 2 ).

INEX is organised jointly under the Evaluation Forum of the DELOS Network of Excellence on Digital Libraries[35] by Norbert Fuhr (University of Duisburg, Germany) and Mounia Lalmas (Queen Mary University of London, UK), and provides an infrastructure for the collaborative creation of a test collection by organisations who wish to participate. Documents for the test collection are provided by INEX, but requests, retrieval runs and relevance assessments are created in a joint effort by the participating organisations. As such the initiative is similar in organisation to, e.g., the Cross Language Evaluation Forum[36] (CLEF), but different from the Text REtrival Conferences[37] (TREC) where documents, requests as well as relevance assessments are provided by the organisers. INEX 2002[38] was completed by the distribution of the final test collection on 30 March 2003, and INEX 2003[39] is currently running. 36 groups from four different continents contributed actively to INEX 2002. Overviews of INEX2002 are given in Fuhr et al. (2002), in Gövert and Kazai (2003), and in the full proceedings of the INEX 2002 workshop (Fuhr et al., 2003).

---

[35] See http://delos-noe.iei.pi.cnr.it/activities/5forums.html

[36] See http://www.clef-campaign.org/

[37] See http://trec.nist.gov/

[38] http://qmir.dcs.qmul.ac.uk/inex/

[39] http://inex.is.informatik.uni-duisburg.de:2003/

INEX is interesting in relation to testing the boomerang effect for a number of reasons:

- It is the first IR initiative that bases its test collection on a reasonably sized corpus of scientific documents in full text including the reference lists.
- The document corpus is marked up in XML which to some extent facilitates the extraction of functional representations from the documents. This type of documents is normally very hard to get hold of in large numbers.
- In addition to the document corpus both requests and relevance assessments are produced in a collaborative effort. This is an important advantage, not only in terms of resources. As the corpus consists of scientific documents from a particular domain, experts are needed to construct requests and assess the retrieved documents if the results are to be reliable. It is unlikely that students or test persons without specific domain knowledge will be able to create realistic requests, or to make realistic relevance assessments as experienced by Shaw, Wood and Tibbo (1991)[40]. In INEX the requests and assessments are provided by domain experts from many different institutions across the globe.
- Similar to TREC, there is a greater possibility of obtaining a more varied and realistic recall base because of the wide range of participants with their different approaches to retrieval. Furthermore, by joining INEX and submitting runs, it can be ensured that the type of documents retrieved by the boomerang effect is assessed and included in the recall base. This is an issue in the main experiment as the boomerang effect might tend to retrieve different documents through the citation network.
- INEX provides a realistic benchmark for the best match boomerang effect, including tools to calculate performance measures that can be compared to the performance of the IR techniques tested by the other participants.

In view of the advantages mentioned above, the INEX 2002 test collection was chosen for the main experiment in the present dissertation as the best available alternative for testing the best match boomerang effect. Access to the test collection is restricted to active INEX participants, and it was therefore decided that the TAPIR project would join INEX 2002[41] with the purpose of testing the best match boomerang effect.

The retrieval task to be carried out in INEX is the ad-hoc retrieval of XML documents:

---

[40] In response to the same requests an information specialist without such knowledge provided markedly different assessments compared to three groups of domain experts.

[41] Referred to simply as INEX in the remainder of the dissertation.

> "Just as in TREC, the ad-hoc task was defined with the aim to evaluate the performance of systems that search a static set of documents using a new set of topics. This task has been described as a simulation of how a library might be used, where the collection of documents is known, while the queries to be asked are unknown [(Voorhees and Harman, 2002)]." (Govert and Kazai, 2003, p. 2 ).

Compared to TREC the main difference is that *document components* can be specified in the requests. A document component is defined as any element of the XML mark-up and can be, for instance, the author names, a sub-section or a figure caption. This makes it possible to create more complex queries and more diversified answer elements. Two types of topics[42] are included in the INEX ad-hoc task to take advantage of this:

- *Content-and-structure (CAS) topics*, which contain explicit references to the XML structure, for instance by confining the search keys or the answer elements to certain document components.
- *Content-only (CO) topics*, which disregard the XML structure similar to the requests used in TREC. "Their resemblance to traditional IR queries is, however, only in their appearance. They pose a challenge to XML retrieval in that the results to such queries can be (possibly overlapping) XML elements of varying granularity that fulfil the query." (Govert and Kazai, 2003, p. 2 ).

The reason that CO requests are included is that users may not know the structure of the documents, and therefore systems should also be able to provide an answer to queries without structural constraints. This resembles the passage retrieval tested, e.g., by Salton, Allan and Buckley (1993), but with the added potential benefit that better passages might be extracted from the document structure as defined by the XML mark-up. In spite of the fact that the INEX organisers place equal weight on both types of topics, it is not unfair to say that most participants in INEX probably regard the CAS requests as the main challenge[43]. Only the CO topics will be used for testing the boomerang effect in the main experiment, however, because only whole documents can be retrieved with the best match boomerang effect in its present form. Whole documents

---

[42] An INEX topic is similar to the ones used in the ad-hoc task in TREC. They consist of a short *title*, a natural language *description,* and an extended *narrative*, and *keywords* that are useful for searching.

[43] It was for instance only possible to complete assessments for 24 of 30 CO requests, while assessors could be found for all 30 CAS requests.

are possible answer elements to all CO topics, but not to all CAS topics because the structural constraints in the latter often require a specific part of a document or combinations hereof as answer.

### *6.1.1 Document corpus*

The documents corpus, donated by the IEEE Computer Society[44], consists of 12,107 items in full text from the society's 12 magazines and 6 transactions in the period 1995-2002 (see Appendix 2). All documents are marked up in a rich XML structure, originally with the purpose of producing the print versions of the journals. The size of the entire corpus file is 494 MB (402 MB excluding the XML tags), and it includes every item in the published journals except for advertisements. Thus, in terms of the included types of documents it is a very realistic collection as, e.g., editorials, news items, letters to the editor, tables of contents, lists of reviewers etc. form part of the corpus along with the actual feature articles and research articles[45]. There are two main types of journals in the corpus: *the IEEE Magazines*, with application-oriented content; and *the IEEE Transactions* with research-oriented publications. Appendix 2 shows the distribution of documents on journals and their file size as well as the journal impact factors of the journals. Table 6.1 below summarises the differences between magazines and transactions. It may be observed that there are fewer, but longer documents in the transactions (35% of the articles, but 60% of the file size). This is due to the fact that the transactions mainly contain research articles, whereas the magazines contain many shorter articles dealing with, e.g., trends and news, and consequently fewer research articles.

**Table 6.1. The distribution of articles on magazines and transactions in the INEX test collection.**

| | Size (MB) | No. of documents |
|---|---|---|
| Magazines | 197 (39.8%) | 7,874 (65.0%) |
| Transactions | 297 (60.2%) | 4,233 (35.0%) |
| | 494 (100.0%) | 12,107 (100.0%) |

[44] See http://www.computer.org/

[45] In the remainder of the dissertation the term *article* will be used to denote all the 12,107 items unless otherwise specified.

The corpus covers several aspects of the computer science field. Because of the association with the IEEE (Institute of Electrical and Electronics Engineers) about a third of the journals mainly deal with the engineering aspects of computer science, e.g., test and design of computers or parallel and distributed systems.

All documents in the corpus are tagged in XML, which conform to one common Document Type Definition (DTD). The overall structure of a typical document is shown in Figure 6.1. It consists of front matter (`<fm>`), with bibliographical data such as author, title, publication data, abstract etc., the text body (`<bdy>`) with the main content of the document, and back matter (`<bm>`) with the bibliography and information about the authors of the document. The text body is structured in sections (`<sec>`), sub-sections (`<ss1>`), and sub-sub-sections (`<ss2>`), and each of these units usually begin with a section title (`<st>`) followed by a number of paragraphs (`<p>`). In addition there is mark-up for internal references (e.g., from the body text to figures, tables or the bibliography), bulleted lists, layout emphasis (e.g., bold, italics or special capital letters), footnotes, etc. The references in the bibliography are marked up so that the following information can be distinguished if available: surname and first name of all cited authors, cited article title, cited journal/work, cited volume and issue, cited pages, and cited year. Examples of references in original mark-up are given in Appendix 3.

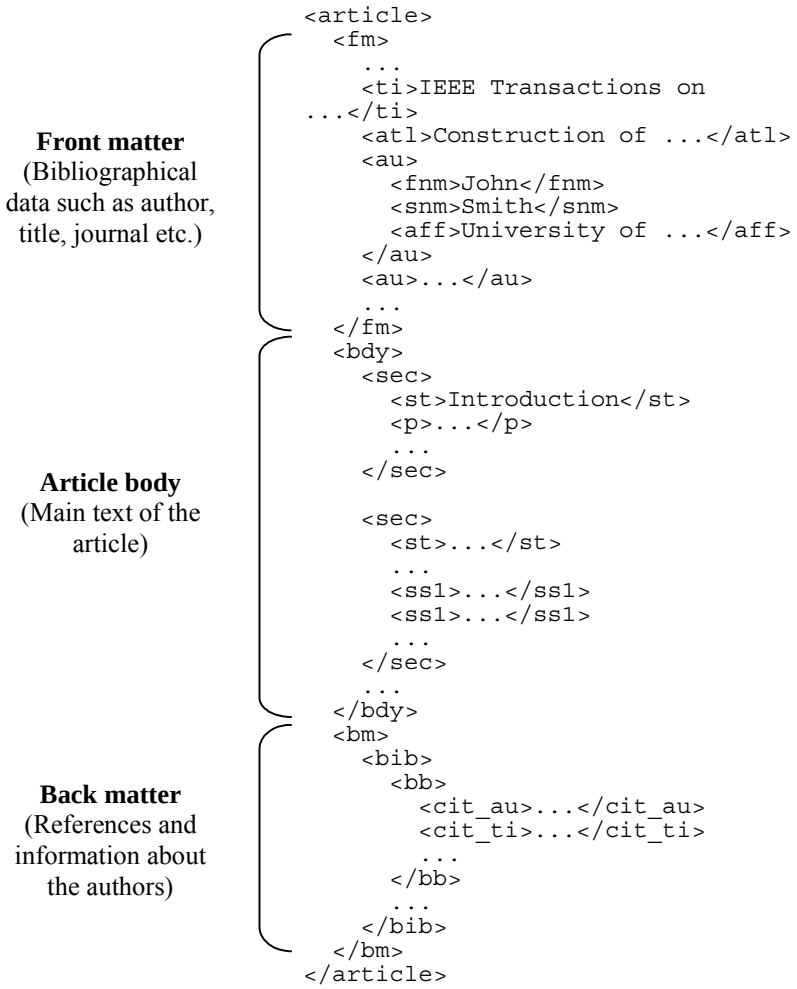



**Figure 6.1. Outline of a typical article in the INEX test collection. Modified from Gövert and Kazai (2003).**

### *6.1.2 Topics*

The topic format in INEX is modified from the format used for TREC's ad-hoc task (Gövert and Kazai, 2003). It consists of four elements: title, description, narrative and keywords. As the CAS topics are not used in the present dissertation, only details related to the CO works are described. An example of a topic is given in Figure 6.2 below, and further three INEX2002 topics are given in Appendix 10. The *title* is a condensed version of the description, and consists of a few keywords that best describe the user's information need. The *description* is a short description of the information need in the form of a few natural language sentences. The *narrative* is a detailed explanation of the information need and gives guidelines as to what makes a document

relevant, or alternatively not relevant. The *keywords* is a list of search keys that have been useful as search keys during the topic creation, and often include synonyms to the words in the title.

```
<INEX-Topic topic-id="39" query-type="CO">
    <Title>
      <cw>Video on demand</cw>
    </Title>
    <Description>
      What is the typical design or architecture of video systems providing
      video on demand (VOD)?
    </Description>
    <Narrative>
      To be relevant, a document/component should contain information about the design
      or architecture of a video server or video processing system providing video
      on demand. A document/component is not relevant if it provides very specific
      technical information like disk scheduling or communication protocols.
    </Narrative>
    <Keywords>video, video on demand, architecture, server, design</Keywords>
</INEX-Topic>
```

**Figure 6.2. Example of a CO topic from the INEX2002 test collection.**

The topics were developed by the participants using guidelines provided by the INEX organisers (See the INEX Guidelines for Topic Development in Fuhr et al., 2003, p. 178-181). The aim of the developments process was to create a diverse set of topics that would retrieve not too many or too few documents in order for the topics to function as a useful diagnostic tool. The author of a topic should preferably be an expert or familiar with the subject area covered by the corpus and should, apart from creating the topic, also be the one to assess the retrieved documents for relevance (see Section 6.1.3 below). The participants were instructed to create topics that reflect what real users of operational systems might ask, and what an operational system might provide. Candidate topics were created in a three-stage process: 1. Creation of initial topic descriptions; 2. collection exploration; and 3. topic refinement. In the first stage an initial topic description was created without regard to specific system capabilities or collection particularities as a short natural language account of a user's information need (this became the description element of the topic after refinement). In the second stage the corpus was explored using any available IR system, and the top 25 documents assessed. Following a refinement of the search statements the top 100 documents were assessed. As far as possible the assessments were to be carried out consistently and without regards to previously seen documents. Any useful keywords used in the search statements were recorded to the keywords element of the topic. In the third stage the narrative was created with a detailed explanation of the information need including

directions as to what would or would not be considered relevant, based on the experiences in the first two stages. Finally, all elements in the candidate topic were refined to ensure consistency and that each element could be used on its own if needed, e.g., for title or keyword searches only. 143 candidate topics were submitted (CO and CAS) and 30 of each were included in the final set of topics by the INEX organisers:

> "The selection of the final 60 topics was based on the combination of criteria, such as…having topics that are representative of IR, DB and XML-specific search situations, balancing the load across participants for relevance assessments, and eliminating topics that were considered too ambiguous or too difficult to judge. We also aimed to include topics that were likely to retrieve diverse sets (varying granularity) of relevant components. Furthermore, we based topic selection on the estimated number of relevant components, where we selected topics with at least 2, but no more than 20 relevant items in the top 25 retrieved components." (Gövert and Kazai, 2003, p. 7).

Examples of the final topics are given in Figure 6.2 (topic 39) and in Appendix 10 (topics 36, 47 and 58). As discussed above it is an advantage that the topics are constructed by a very diverse group of experts from many different institutions with many different research interests. This maximises the likelihood that a broad range of subjects are included, and reduces the risk of biases that might arise from creating topics at a single institution. Because the topic authors are mainly self selected by their interest in joining INEX there is a bias towards topics that in some way deal with IR (9 out of the 25 assessed CO topics)., and probably fewer topics that deal with the engineering aspects of computer science that would be expected taking the nature of the corpus into consideration. The main disadvantage is that topic authors with many different backgrounds have to understand and follow the topics development guidelines. Problems with different interpretations may arise, for instance some topic authors have indicated phrases in the topic keywords (e.g., topic 39 and 47), while others have not (e.g., topics 36 and 58).

An interesting observation may be made considering the topic development guidelines and the way they have been interpreted as manifested in the final topics: the topics elements may to a certain extent be mapped onto the three functionally different contexts that might be extracted from the user's cognitive space, as proposed in Ingwersen (1996, p. 18ff). Terms in the title and keywords elements can be regarded as different *request versions*. The keywords have certainly been accumulated in this way.

Similarly, the description and narrative elements can be regarded as more or less elaborate *problem or goal statements*. However, nothing that resembles the *work task or interest description* is included explicitly in neither the topic development guide nor the final topics. The partial correspondence is interesting though, as the topic format could be modified to include the work task or interest description as well, e.g., by adding the requirement in the guidelines that the underlying reasons for having the information need (the 'why') should be given in the narrative. If such modifications were made, the INEX topics could also be presented as simulated work task situations to other test persons than the topic author, as proposed by Borlund (2000a; 2000b). Thereby the value of the INEX test collection would be enhanced as it could be used not only in laboratory experiments as intended, but also in interactive experiments involving users. Comparisons between the two types of experiments would then be possible with the same set of work tasks.

### *6.1.3 Relevance assessments*

For each topic a recall base was generated by pooling the results from the retrieval runs submitted by the participants. A maximum of three runs was permitted per participant, and 49 different CO runs were submitted. Each run consisted of the top 100 documents or components retrieved by the IR system used (That is, the document cut-off value (DCV) was 100). The submitted result elements varied from authors, titles and paragraphs over sections and sub-sections to whole articles and even journals. The submissions were merged into a pool for each topic as proposed in Sparck Jones and van Rijsbergen (1976) and as used, e.g., in TREC. A median sized CO pool contained 1980 document components from 981 documents, and there was a pooling effect, i.e., a reduction in pool size because of overlapping items, of 69% on the document level in the CO topics. A total of 30,275 documents, and 60,066 components were included in the CO pools (see Table 6.2 below). All documents and components in a pool were returned to the topics authors for relevance assessments. The person assessing a pool should preferably be the same as the topic author. However, a number of participants dropped out because of resource problems, and ten CO topics were not assessed by their original authors. For four of these topics, volunteers with subject expertise were found, and a total of 24 CO topics with relevance assessments are included in the final INEX test collection.

The relevance of the answer elements in relation to the specification given in the topics were assessed along the following two dimensions (Gövert and Kazai, 2003, p. 8):

- "**Topical relevance,** which reflects the extent to which the information contained in a document component satisfies the information need.
- **Component coverage,** which reflects the extent to which a document component is focused on the information need, while being an informative unit."

In spite of the name the topical relevance dimension in INEX should not be confused with the type of relevance that Saracevic calls "a topical-like type, associated with aboutness" (1996). This is the type of relevance judgements made by human assessors in IR experiments like TREC, where concept of topic is understood as aboutness (Hutchins, 1978). Borlund names this type of relevance *intellectual topicality* to emphasise that the assessment is not objective but rather "…an intellectual assessment of how an information object corresponds to the topical area required and described by the request(s) for information." (Borlund, 2003, p. 915). In accordance with Borlund the actual relevance type applied in INEX rather belongs to either pertinence or situational relevance along Borlund's continuum of subjective relevances, since the CAS topics may also deal with properties of documents *other* than contents, such as author names. *Pertinence* is commonly defined as the relationship between information need and retrieved objects as perceived (Saracevic, 1996; Cosijn and Ingwersen, 2000) – exactly the definition applied to the topical relevance dimension by Gövert and Kazai. *Situational relevance* is regarded as potentially dynamic by Borlund and "…expresses the relationship between the user's perception of usefulness of a retrieved information object, and a specific work task situation." (2003, p. 922). Since the topic authors were also meant to be the relevance assessors, situational factors can easily have affected the assessments, because the topic authors can cognitively know more about the underlying work task that is behind the information need as expressed in the INEX topics. In the instructions given to assessors in INEX steps were taken to reduce the influence of situational factors on the assessments: "You should judge each document component on its own merits. That is, a document component is still relevant if it is the twentieth you have seen with the same information!" (From the INEX Relevance Assessment Guide, p. 186 in Fuhr et al., 2003). As in any pursuit involving human interpretation there is a risk in INEX that these instructions were not followed entirely consistently, and it can be difficult in practice to assert whether the actual relevance type applied is pertinence or situational relevance as noted by Borlund (2003). Regardless of the relevance type, the main strength of INEX is, in our view, that the scientific documents in the corpus

are judged by assessors that are domain experts in relation to information needs that they have formulated themselves.

The *degree of relevance* (See, e.g., Saracevic, 1996) between the information need as expressed in the topic and the retrieved answer elements was assessed using graded relevance for both dimensions. Figure 6.3 shows the scale for topical relevance assessments used in INEX. It is measured on a four-point ordinal scale adapted for INEX from one used in the FIRE[46] laboratory at Tampere University, Finland (see e.g., Sormunen, 2002).

- **Irrelevant (0):** The document component does not contain any information about the topic of request.
- **Marginally relevant (1):** The document component mentions the topic of request, but only in passing.
- **Fairly relevant (2):** The document component contains more information than the topic description, but this information is not exhaustive. In the case of multi-faceted topics, only some of the sub-themes or viewpoints are discussed.
- **Highly relevant (3):** The document component discusses the topic of request exhaustively. In the case of multifaceted topics, all or most sub-themes or viewpoints are discussed.

**Figure 6.3. The four-point ordinal scale for the topical relevance assessments used in INEX2002. (Gövert and Kazai, 2003, p. 8).**

The main difference is that specifications of the typical extents of relevant material in documents given in Sormunen (2002) have been removed and are measured separately in the assessment of component coverage in INEX. The measure of component coverage is particularly important in XML IR where document components can be returned as answer elements. The component coverage was measured on a four-point nominal scale adapted from Schütz (1998):

[46] The Finnish Information Retrieval Experts Group (http://www.info.uta.fi/tutkimus/fire/).

- **No coverage (N)**: The topic or an aspect of the topic is not a theme of the document component.
- **Too large (L)**: The topic or an aspect of the topic is only a minor theme of the document component.
- **Too small (S)**: The topic or an aspect of the topic is the main or only theme of the document component, but the component is too small to act as a meaningful unit of information.
- **Exact coverage (E)**: The topic or an aspect of the topic is the main or only theme of the document component, and the component acts as a meaningful unit of information.

**Figure 6.4. The four-point nominal scale for assessments of component coverage used in INEX2002. (Gövert and Kazai, 2003, p. 9).**

As noted by the INEX organisers the two relevance dimensions "…are not perfectly orthogonal to each other. Some combinations of relevance / coverage values do not make sense…" (Gövert and Kazai, 2003, p. 9). For instance an irrelevant document component (0) cannot have any coverage, and a document component that is too small (S) cannot be highly relevant (3). The allowed combinations were: 3E, 2E, 1E, 3L, 2L, 1L, 2S, 1S and 0N. Note that none of the scales are at interval or ratio level, and that you cannot infer, e.g., that a highly relevant document is three times better than a marginally relevant one. The assessments were recorded in an online assessment system, which checked the formal consistency of assessments (see Gövert and Kazai (2003) for details and screenshots of the system). The most frequently returned relevant document components (topical relevance > 0) were: paragraph (`<p>`), section (`<sec>`), body (`<bdy>`), sub-section (`<ss1>`), sub-sub-section (`<ss2>`), and title (`<atl>`) (Gövert and Kazai, 2003). Table 6.2 shows the distribution of the relevance combinations on the level of whole documents and components (excluding whole documents). As could be expected a considerable proportion (69%) of the whole documents were assessed as being too large (L). At the component level there was an almost equal proportion of too large (L) components (38%) and too small (S) components (43%). As most participants returned document components as answer elements, there were of course many more of these compared to whole documents, but it is interesting to note that the proportion of relevant documents with exact coverage (E) is nearly the same at the document and component levels (22% and 19% respectively).

**Table 6.2. Distribution on relevance combinations of the collected relevance assessments for the CO topics (topical relevance > 0). The component column excludes whole documents. Modified from Gövert and Kazai (2003)**

| Relevance assessment | Document level | Component level | Document level | Component level |
|---|---|---|---|---|
| 3E | 307 | 1,087 | | |
| 2E | 165 | 1,107 | 22.0% | 18.7% |
| 1E | 114 | 827 | | |
| 3L | 394 | 1,145 | | |
| 2L | 599 | 2,295 | 69.3% | 38.1% |
| 1L | 854 | 2,708 | | |
| 2S | 118 | 3,825 | 8.8% | 43.2% |
| 1S | 116 | 3,156 | | |
| All relevant | 2,667 | 16,150 | 100.0% | 100.0% |
| All retrieved | 30,275 | 60,006 | | |

Two sets of relevance assessments are available in the INEX test collection as basis for the calculation of performance measures: the *original* set of assessments as recorded, and a set containing *implicit* assessments automatically generated from the former. The implicit assessments cover documents and components that have not been assessed explicitly, but for which it is nonetheless possible to deduce assessments from the document structure. This is possible in two cases (Gövert and Kazai, 2003, p. 11):

- "Due to the definition of the relevance dimension, the relevance level of a parent component of an assessed component is equal to or greater than the relevance of the assessed component.
- For a component which has a coverage assessment of *exact* or *too large* it can be deduced that its parent component has a coverage of *too large*."

The implicit assessments represent an *optimistic* relevance propagation where a document is considered relevant if just some of its contained nodes are assessed as relevant (Roelleke et al., 2002; Kazai, Lalmas and Reid, 2003). A pessimistic relevance propagation would, e.g., require that all of the nodes of a document have to be relevant in order for the document itself to be considered relevant. In terms of the INEX test collection, the optimistic relevance propagation results in an enlargement of the recall base. For instance, the number of documents that are considered as highly relevant with

exact coverage (3E) is expanded from 307 documents (see Table 6.2) to 1396 documents (see Appendix 6) when implicit assessments are used. The realism of this in all situations in relation to CO topics can be disputed, e.g., if a paragraph has been deemed to be fairly relevant but too small (2S) the whole document will be considered as fairly relevant (2) if it is retrieved and no other components of the document have been assessed. While this can be seen as a problem it can be eliminated by the scoring functions used when calculating performance measures (See Section 6.4.1 below).

### *6.1.4 Testing the boomerang effect in INEX*

As mentioned above INEX provides a unique opportunity to test the best match boomerang effect because the corpus consists of scientific full text articles including the references, and because the topics and relevance assessments were created by experts in an international collaborative effort. It is important to note that without the CO topics, the boomerang effect could not have been tested in INEX in its present form as it retrieves whole documents, not parts of them. If the INEX organisers had decided to include CAS topics only the topics and relevance assessment could not have been used, and would have had to be created for the testing of the boomerang effect. Even with the CO topics a matter of concern is whether the recall base will be adequate as a baseline since too few whole documents might be retrieved (and therefore assessed) by the other IR approaches that mainly retrieve document components. It was therefore decided to submit two baseline runs in addition to a best match boomerang effect run as part of the TAPIR project's participation in INEX to ensure an adequate baseline. A different matter of concern in INEX is that although a reasonably large number of documents are included in the corpus, the relevant documents for most topics might tend to occur in one or a few journals only because of specialisation of the individual journals. This will affect all participants in INEX, but might be severe for an approach that retrieves whole documents only because much fewer answer elements are possible (See Table 6.2). Neither concerns turned out to be a problem that will affect the best match boomerang effect more than other systems because it was decided to use upward relevance propagation in INEX (See above and Section 6.4.1 below). As a consequence a whole document will be considered relevant if any component in it is assessed as relevant. Thereby the baselines turned out to be partially unnecessary for eliminating these concerns. The baselines did, however, provide other interesting results (See Section 6.3.3 and Section 7.1 below).

## 6.2 IR system and test database

As no IR facilities capable of handling the necessary operations were available, it was decided to initiate the establishment of an IR laboratory at the Royal School of Library and Information Science in Copenhagen[47]. The *InQuery* system (version 3.2 compiled for Linux Red Hat 6.1) was chosen for the test of the best match boomerang effect. One reason was that InQuery provides the possibility of weighting the search keys, which is needed for Step 3 in the best match boomerang effect. On a long view another motive is that InQuery seems appropriate for other tests of the theory of polyrepresentation as the ideas behind InQuery have inspired the conception of the document space in the theory of polyrepresentation (see, e.g., Figure 2, p. 107 in Ingwersen, 1994). In addition InQuery has performed well in a number of experiments, for instance in TREC (Sparck Jones, 1998), and has found practical use as an operational search engine[48]. However, any best match system could have been used, provided that it offers the possibility of weighting the search keys.

### *6.2.1 The InQuery IR system*

InQuery has been developed at the Center for Intelligent Information Retrieval (CIIR), Computer Science Department at University of Massachusetts. It is based on a particular form of the probabilistic IR model called *the inference network model* (Turtle and Croft, 1990; Turtle, 1991). An advantage of the model is that it can represent many IR approaches and combine them in a single framework. For instance, Turtle and Croft (1991) show how the exact-match and vector space models can be described within the inference network model. Thereby experiments with combinations or comparisons of several models can be conducted in the same framework. The basic idea in the inference network 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). Figure 6.5 shows a basic document retrieval inference network. The

---

[47] A Linux server and a laboratory assistant were partly funded by the TAPIR project and partly by the existing IT lab under the Department of Information Studies.

[48] A number of US governmental full text collections are for instance accessible via InQuery (See http://ciir.cs.umass.edu/demonstrations/membersitedemos.html).

network has two components: a document network and a query network. The document network consists of the documents as abstract entities, which form the root nodes of the network ($d_1$ ... $d_i$), their physical representations as texts ($t_1$ ... $t_j$), and different types of representation concepts derived from these ($r_1$ ... $r_k$). If there are several different versions or formats of the same document these may be represented by different texts ($t_1$ ... $t_j$). Usually there is only one version in experimental IR, resulting in a one-to-one correspondence between documents and texts as shown in Figure 6.5. This is also the case in the INEX corpus, but other cases may arise where there will be different versions of the same document, for instance on the web where a pre-print of a paper may appear in several versions along with the final published version. The document network is constructed once for a given collection at indexing time, and does not change when queries are processed. The query network consists of the user's information need ($I$) expressed as one or more queries ($q_1$ ... $q_n$), as well as query concepts ($c_1$ …$c_m$) derived from the queries. The query network is constructed for each information need at run-time, and may be changed during retrieval, for instance if relevance feedback is carried out.

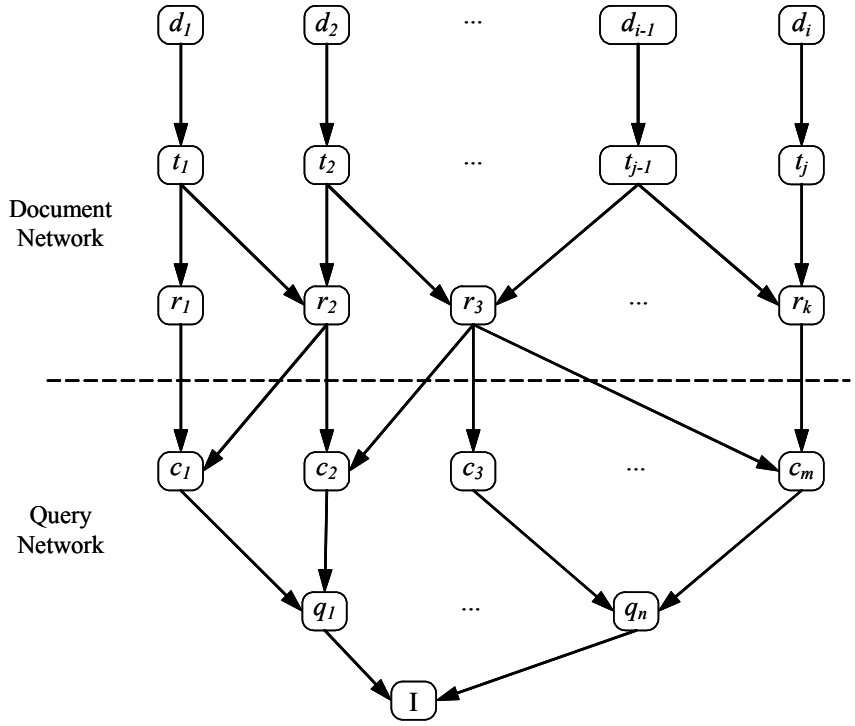



**Figure 6.5. Basic document inference network (Modified from Turtle and Croft (1990, p. 4), and Kekäläinen (1999, p. 20)).**

At run-time the document and query networks are joined through arcs between the representation concepts derived from the documents and the query concepts derived from queries. This is where the match takes place in relation to Figure 2.2. The nodes in the network represent propositional variables and the arcs represent probabilistic dependence relations between these propositions. Nodes in the network are either *true* or *false*, but the arcs take values that range from 0 to 1 and are interpreted as belief values, e.g., the belief in the proposition that a user's information need is satisfied by a particular document. The inference network model is based on Bayesian belief networks, which are directed acyclic graphs, that is, the arcs in the graph go in one direction only, from the parent nodes to the child nodes. The direction of the arcs in the inference network for document retrieval as shown in Figure 6.5 may seem counterintuitive. This is, however, a feature of Bayesian belief networks where the relationships expressed by the arcs are *causal*, i.e., the parent nodes cause the child nodes (Baeza-Yates and Ribeiro-Neto, 1999). The strengths of the causal influences between parent and child nodes are expressed by conditional probabilities and constitute a joint probability distribution for the nodes in the network. Thereby IR becomes "…an inference or evidential reasoning process in which we estimate the probability that a user's information need, expressed as one or more queries, is met given a document as "evidence"." (Turtle and Croft, 1990, p. 1)[49]. In retrieval the documents in the network are therefore observed one after one. When a single document $d_i$ is observed, evidence is attached to the network asserting $d_i$ = *true* and all other documents set to *false*. Because $d_i$ is being observed (i.e., $d_i$ = *true*) the belief for every node in the network can now be calculated. In particular, the probability that the information need is met by $d_i$ can be calculated. By repeating this process for all documents the probability that the information need is met given each document in the collection can be computed, and the documents can be ranked accordingly (Turtle and Croft, 1990, p. 47).

Each document node has a prior probability associated with it that describes the likelihood that the document is observed – this is typically set to *1/the collection size*. The probabilities associated with all non-root nodes in the document network and query

[49] This idea is taken to the extreme in applications of Dempster-Schafer's uncertainty theory in IR where the probabilities in the network are used to *prove* the query in a formal mathematical logic proof (See e.g., Ruthven and Lalmas (1998)).

network must be estimated given the particular set of values for its parent nodes. This estimate is encoded in a link matrix, which captures the probabilities that a given node takes on the value *true* or *false* for all combinations of its parent values. For the representation concepts in the document network ($r_1$ ... $r_k$) the link matrix is generated at collection building and incorporates the effect on any indexing weights (for instance the term frequency for each parent text) or term weights (for instance the inverse document frequency) associated with the representation concept. The belief in a representation concept is estimated by the function:

$$\alpha + (1-\alpha) * ntf * idf \qquad (2)$$

where α, known as the default probability, represents the probability that an index key should be assigned to a document in which it does not occur[50], *ntf* the normalised term frequency, and *idf* the inverse document frequency. The precise formulation in InQuery version 3.2 for the estimation of conditional probabilities for the representation concept nodes is (Allan et al., 1997):

$$0.4 + 0.6 * \left( \frac{tf_{ij}}{tf_{ij} + 0.5 + 1.5 * \frac{dl_j}{adl}} \right) * \left( \frac{\log\left( \frac{N + 0.5}{df_i} \right)}{\log(N + 1.0)} \right) \qquad (3)$$

where $tf_{ij}$ = the frequency of the index key $i$ in document $j$
$dl_j$ = the length of document $j$ (i.e., the number of index keys)
$adl$ = the average document length in the collection
$N$ = the collection size
$df_i$ = the number of documents containing index key $i$

The link matrices for the nodes in the query network are built at run-time for each information need. In InQuery different link matrices can be constructed to accommodate

[50] The incorporation of this α-factor has been shown to improve performance, see (Turtle and Croft, 1990; Turtle and Croft, 1992).

different query operators, for instance, the exact match Boolean operators (AND, OR, NOT), 'softer' probabilistic versions of these, as well as the weighted sum operator, which is of special interest to the best match boomerang effect. Many more are possible, but a limited number of canonical link matrices have been implemented to reduce the space requirements (Turtle and Croft, 1991, p. 42). A list of InQuery's basic query operators are given in, e.g., Callan, Croft and Harding (1992) or Rajashekar and Croft (1995), and an extended list with examples can be found at the homepage of FEMA[51]. Only weighted sum (`#wsum`) and the sum (`#sum`) operators are considered here as none of the others are used in the main experiment in the dissertation. The idea behind the `#wsum` operator is that in "…probabilistic retrieval each parent has a weight associated with it, as does the child. In this weighted-sum matrix, our belief in Q [the present node under consideration] depends on the specific parents that are true – parents with larger weights have more influence on our belief." (Turtle and Croft, 1990, p. 9). The weights of the query keys can be specified in a weighted sum query which is defined as follows (the weights of the representation concepts are specified by the function in Formula 3):

$$P_{wsum}\left(w_s, w_1Q_1, w_2Q_2, \cdots, w_qQ_m\right) = \frac{w_s\left(w_1p_1 + w_2p_2 + \cdots + w_qp_m\right)}{\left(w_1 + w_2 + \cdots + w_q\right)} \qquad (4)$$

where $P$ denotes probability, $Q_m$ ($m = 1 \ldots n$) a query key or an InQuery combination of query keys, $p_m$ the belief value of $Q_m$, $w_q$ the weight of $Q_m$ as specified in the weighted sum query, and $w_s$ the weight of the whole clause (if the user wishes to give higher weights to all the query keys). The `#wsum` operator is a soft probabilistic query operator in the sense that if there is more than one query key only one of them have to present in the retrieved documents (unlike, e.g., the exact match Boolean AND operator where all query keys have to be found in the documents). The weights in the `#wsum` operator "…[s]pecifies that some terms are more important than others, but also that the more present, the better." (Rajashekar and Croft, 1995, p. 275). The `#sum` operator, which is invoked as the default operator in InQuery, is a special case of the `#wsum` operator where all the weights are (implicitly) set to 1. If a query is entered in natural language

[51] FEMA, the Federal Emergency Management Agency, uses InQuery to index their collections, see http://www.fema.gov/search/advsrch.shtm

into InQuery without any operators the default operator applied by InQuery is `#sum`, except if one of the query keys occurs more than one time – then the `#wsum` operator is applied to express this. See Section 6.3 and subsections for the use made of the `#sum` and `#wsum` operators in the dissertation. Figure 6.7 in Section 6.3.1 gives an example of the transformation of InQuery's parsing of queries.

### *6.2.2 Representations generated from the INEX corpus*

A number of functional and cognitive representations were generated for use in the main experiment. These are shown in Table 6.3 below together with associated statistics. The functional representations were generated directly from the INEX corpus and are described in this section, with the exception of the citation code and citation index, which are discussed separately in Section 6.2.4. Cognitive representations generated from external sources are discussed in Section 6.2.3. Although some of the document types are unlikely to be relevant to any of the CO topics, e.g., tables of contents or lists of reviewers, it was chosen not to attempt a removal of any of these in order to make the retrieval task as realistic as possible.

The functional representations were extracted from the XML structure of the documents in the corpus using XSL transformations[52]. The mark-up was originally made primarily to facilitate the production of the printed versions of the journals. This means that every component that might need to be typeset in a different font or with different layout has been marked up with XML tags. The goal of the extraction for the main experiment has been to utilise this elaborate syntactic mark-up to extract representations that will provide meaningful functional representations of the documents. As scientific publications are one of the examples used to illustrate the theory of polyrepresentation, a number of the explicitly addressed representations have been chosen for the main experiment (see e.g., figure 7, p. 32 in Ingwersen, 1996): The article title, figure captions, table captions, and cited titles. It is possible to extract many more, but for want of time to design extraction filters only a few more functional representations were extracted: The article abstract, author keywords if provided, as well as the introduction and conclusion sections of the articles. One notable representation recommended for

[52] The eXtensible Style Sheet Language (XSL) allows transformations of XML documents into new formats – see http://www.w3.org/TR/xsl

testing in Ingwersen (1996) that could not be extracted directly is the headings from sections and sub-sections. The principle of using headings as well as table and figure captions for document representation was proposed and tested in the Subject Access Project (SAP) where they were utilised successfully for monographic literature (Atherton-Cochrane, 1978; Wormell, 1985). The article title and abstract have been used for document representation in numerous studies and operational systems, and are intuitively strong semantic representations as they function as a summary of the whole document, especially in a field such as computer science where many articles are either technical or theoretical. H. P. Luhn's famous KeyWords-In-Context (KWIC) indexes based on words extracted automatically from titles provide strong support for their use as a separate functional representation (Luhn, 1960). The author keywords are less strong as functional representations as there are weaker conventions concerning how they should be selected, but they serve the same summarising purpose as the title and abstract. The cited titles are included because they are interesting from a cognitive point of view as discussed in Chapter 3: the references are chosen by the author, but follow the conventions of the field to a certain extent (Small, 1978). The actual words of the cited titles have, however, been chosen by the *cited* author, and are beyond the influence of the author of the article. Note that the cited titles in the INEX corpus are more complete than the ones used in the work of Kwok (e.g., 1975; 1985b), Salton and Zhang (1986), and in ISI's KeyWords Plus (Garfield, 1990) as *all* cited titles are available in the INEX corpus. In the applications mentioned only those cited titles that are included as source items in the used citation index could be utilised. See Section 6.2.4 for further details of the citation data available in the INEX corpus.

**Table 6.3. Document representations extracted for the main experiment. The number of documents, number of unique index keys, and the total number of index keys are derived from InQuery, i.e., they do not include stop words and have been stemmed (See Section 6.2.5). The size is the number of MB of the representations before indexing, but excluding tags and ID numbers.**

| Type | Name (Abbreviation) | No. of documents | No. of unique index keys | Total no. of index keys | Size in MB |
|---|---|---|---|---|---|
| Functional | Article title (ATL) | 12,077 | 6,653 | 63,353 | 0.6 |
| Functional | Abstract (ABS) | 7,358 | 13,687 | 452,866 | 5.1 |
| Functional | Author keywords (AKW) | 3,768 | 3,664 | 53,421 | 0.5 |
| Functional | Figure captions (FGC) | 8,159 | 32,923 | 925,538 | 8.5 |
| Functional | Table captions (TBC) | 3,836 | 8,028 | 92,521 | 0.8 |
| Functional | Introductions (INT) | 3,801 | 19,699 | 1,407,503 | 16.7 |
| Functional | Conclusions (CON) | 2,835 | 11,371 | 424,534 | 5.0 |
| Funct. / Cogn. | Cited titles (CTI) | 7,213 | 21,598 | 755,178 | 7.7 |
| Cognitive | Descriptors (DE) | 7,711 | 1,427 | 57,495 | 0.5 |
| Cognitive | Identifiers (ID) | 7,711 | 9,172 | 166,423 | 1.5 |
| Citation Index | CI / flat (f) | 7,111 | 70,634 | 111,741 | 1.4 |
| Citation Index | CI / expanded (x) | 7,111 | 70,634 | 192,881 | 2.5 |

The introduction and conclusion sections of the articles are intuitively appealing as document representations. Ingwersen recommends that "…the functional difference between the locations of sections in the text should be taken into account." (Ingwersen, 1996, p. 33). From the typical structure of scientific articles the introduction, methodology, analysis, discussion and conclusion sections offer such different functional characteristics at different locations in the body of the article text (See, e.g., Bazerman, 1988; Swales, 1990). Of these the most interesting can be said to be the introduction and discussion sections. The introductions are generally used to delimit the area of investigation, review earlier contributions and state the purpose of the article, and the discussion elaborates on the reported results and their consequences (Bazerman, 1988). In a field like computer science the conclusion is mainly used to sum up the whole article, and is therefore similar in function to the abstract. Unfortunately, although the sections themselves are tagged there is no indication as to the type of content in the section. Specialised parsers were written in an attempt to extract the sections based on their content. A special switch in the XSL specification, which allows

the extraction of the first and last elements of a given component, could be utilised to extract the first and last sections in the text body. After initial tests this was found to work well for extracting introductions and conclusions from articles in the IEEE Transactions, but not so well on articles from the IEEE Magazines. The reason for this is that IEEE Transactions mainly contain research papers, whereas the IEEE Magazines have a large proportion of other types of articles that do not conform to the conventions found in research articles. After adding a few simple heuristics[53] it was decided only to include the introduction and conclusion representations for the IEEE Transactions. The extraction and use of introductions and conclusions as document representations are supported by Lahtinen. He found that the first sentence of the first paragraph and the last sentence of the last paragraph of a small test corpus of articles contained a significantly higher proportion of good index terms (Lahtinen, 2000, p. 139-143). Although whole introductions and conclusions were extracted for the main experiment, this should still provide an advantage. The rest of the full text was not used in the main experiment as it proved to be very difficult to extract this consistently without including the representations already extracted. In order to avoid a double indexing of the functional representations with the potential danger of eliminating the desired effect of these, it was decided not to use the whole full text as a representation. Approximately 11 % of the 402 MB text in the INEX corpus (excluding tags) was included as functional representations in the main experiment.

Since each journal in the corpus contains some amount of material that is not likely to satisfy any information needs (for instance annual lists of reviewers, call for participations etc.) it was considered if some of these should be excluded. This was not done in order to keep the task of retrieving relevant documents as realistic as possible. It can be observed from Table 6.3 that the extent to which the documents in the corpus include the functional representations is very varied. Titles are included in almost all documents (99.7%)[54], but the other representations occur in much fewer documents: Abstracts are included in 61%, figure captions in 67%, cited titles in 60% and table captions in only 32%. These representations occur fairly evenly across the journals.

[53] The first and last sections of the body text were extracted and if the words Introduction or Conclusion appeared in the first section title of these the sections were included.

[54] There are a few items of editorial nature that do not have a title marked up in the XML.

Author keywords are only present in the Transactions (31% of all documents and 89% of the Transactions), as are the introductions and conclusions because of the problems with defining a parser to extract these from the Magazines. Introductions occur in 31% of all documents and in 90% of the Transactions. The corresponding figures for conclusions are 23% and 67%. This is what can be expected when the theory of polyrepresentation is applied on a heterogeneous collection of documents. With the present selection of representations and without the remaining full text the consequence is that some of the documents have a larger probability of being retrieved, and thereby form part of overlaps, because they are indexed with more representations than other documents. Although a similar problem may arise when short documents in a full text collection are indexed in a traditional best match system, the effect is levelled out by incorporating normalisation for document length in the TF weighting (See Formula 3 above). The consequence of the unevenness as to the number of representations per document in the main experiment is that research papers have better chances of being retrieved, although none of the other document types are filtered out explicitly. Papers from the Transactions, which are mainly research papers, are the only ones to be indexed with author keywords, introductions and conclusions. The rest of the representations in Table 6.3 tend to occur in most of the research papers as well. In relation to the INEX CO topics this is probably an advantage as the topics tend to be research oriented. In an operational system this might not be desirable in every case. The possibility of giving preference to certain document types could be offered to alleviate this.

### *6.2.3 Cognitive representations*

As the INEX corpus only contains author generated representations, cognitive interpretations of the documents by other agents had to be obtained from other sources. The index terms assigned intellectually in the IEE INSPEC database[55] were chosen as cognitive representations for the main experiment. INSPEC covers scientific and technical literature in physics, electronics and computing including all the journals in the INEX corpus, and provides a high quality intellectual indexing carried out by domain experts *as indexers*. The documents in the INSPEC database are indexed in a number of different ways, including e.g., indexing of numerical and chemical

[55] See http://www.iee.org/Publish/Inspec/

information when necessary. The two main textual indexing fields, *descriptors* and *identifiers*, were chosen for the main experiment as they contain information that can be matched directly with the INEX queries. The descriptors in INSPEC are chosen from the INSPEC Thesaurus. The identifiers are uncontrolled words and phrases chosen freely by the indexers. INSPEC does a selective indexing, where articles that are either outside the scope of the database or are considered to be of little value for the users are excluded from the database. Therefore only a total of 8084 articles from the journals in the INEX corpus were indexed in INSPEC. As there were no common unique identifiers, it was necessary to carry out a match to pair the INSPEC records with the INEX documents. Using an exact match it was possible to match 7711 (95%) of the INSPEC records to INEX documents (See Table 6.3). Thereby 64% of the INEX documents are represented by descriptors and identifiers in the main experiment. These are both recommended for testing in the theory of polyrepresentation, and can be considered to be strong alternative cognitive representations of the documents. Both are assigned by domain experts (that is, a different cognitive agent than the author). The rigour of INSPEC's indexing and the high quality thesaurus behind it contribute to the strong cognitive characteristics of the descriptors. Furthermore, although uncontrolled, the identifiers are relatively plentiful when compared to index keys from the article title: there is an average of 5.2 index keys per article generated from the titles (63,353/12,077), and 21.6 index keys per article generated from the identifiers (166,423/7,711) - Table 6.3. A disadvantage with the descriptors is that the controlled and therefore artificial language in the thesaurus can potentially result in few matches with natural language queries. Different forms of query expansion can be applied to alleviate this problem and even improve IR performance (See, e.g., Kekäläinen, 1999).

### *6.2.4 Citation code and citation indexes*

In order to exploit the references and citations in the INEX corpus a citation code had to be established. The primary purpose of the citation code is to facilitate the discovery of citations to the same article, because references routinely contain errors and inconsistencies as discovered, e.g., when constructing CiteSeer (Lawrence, Giles and Bollacker, 1999). Compared to the input data available to CiteSeer, the INEX corpus offers a much more structured format for the references which is used across all the journals. Appendix 3 shows examples of references with the XML mark-up from the INEX corpus. For most references subfields with the following types of content can be distinguished if available: surname (`<snm>`) and first name (`<fnm>` – usually as initials)

of all cited authors (<au>), cited paper title (<atl>), title of the cited journal/proceeding/book (<ti>), cited volume (<volno>) and issue (<issno>), cited page numbers (<pp>), and publication details such as cited month (<mo>) and year (<yr>). Note that the tag for the title of the cited journal/proceeding/monograph is the same (<ti>), and that these three types of cited work cannot be distinguished within this subfield. The main difference in formatting between monographs on one side, and journal articles or conference papers on the other is that monographs do not usually have a cited paper title (<atl>), except when a specific chapter is named in the reference. 7,276 of the document in the corpus contained references – 7,111 of these documents were included in the citation index created with the methods described below corresponding to 58.7% of the 12,107 documents in the corpus.

Many different ways are possible when constructing a citation code from the subfields. The best performing algorithm for CiteSeer was found to be one that disregards the different types of content, and instead does a match on single terms and phrases on the whole reference (Lawrence, Giles and Bollacker, 1999). This was probably due to the fact that no common structure was found in the CiteSeer data. As more structured data is available in the INEX corpus it was decided to base the citation code on information extracted from the subfields. One possible way of doing this would be to build a database of all references that stores the information from all subfields, and then do an exact match on combinations of these fields. However, even though the mark-up is consistent across all documents in the corpus, the contents in the subfields are provided by the article authors. The consequence is that there are often errors, inconsistencies and omissions in all of the subfields. Therefore an exact match on any of the subfields or combinations of them will result in many false negative matches, and many citations to the same document will not be found. Instead of such a match it was chosen to base the citation code on the *cited paper title* (<atl>). Although the cited paper titles also contain errors and inconsistencies they are the most specific subfield of the references. It was chosen to reduce the errors and inconsistencies in the INEX corpus using an approximate string matching program, i.e., a program that allows non-exact matching of text strings. A similar approach was tested in CiteSeer on whole references, but it did not perform as well as the combined word and phrase matching. Part of the reason for this was that whole citations were used and that these could be of significantly different lengths (Lawrence, Giles and Bollacker, 1999, p. 5 in the preprint). The reduction of errors and inconsistencies was accomplished by using the *like* program developed by

the Department of Information Studies, University of Tampere, Finland for testing a novel application of n-grams called skip-grams or s-grams (See, e.g., Pirkola et al., 2002)[56]. The like program has two main components: an index structure which allows efficient matching of strings, and several forms of approximate string matching algorithms including an *edit distance algorithm*[57]. The edit distance is defined as the smallest number of insertions, deletions, and substitutions required for changing one string into another - therefore the greater the edit distance, the more different the strings are. In the like program a string may be matched against the index to identify the strings that have the smallest edit distance to the string up to a given threshold. The edit distance algorithm in the like program does not, however, take the *length* of the string into consideration. Therefore the same absolute edit distance between two strings may be very different for short and long strings. For instance, with an edit distance of 8, an 8 character sting can be changed into any other string of the same length, whereas a 100 character string can only be modified slightly. Hence, in the match of cited paper titles with the like program, it was necessary to weight the edit distance in proportion to the length of the string. A similar approach was taken by Lawrence, Giles and Bollacker (1999).

7,276 documents in the INEX corpus contained a total of 141,631 references (including those duplicates that occur in several documents). 118,191 of these (84 %) had a cited paper title subfield, that is, they are predominantly journal articles, conference papers, or named book chapters. The rest do not have a cited article subfield, and are mainly citations to monographs. These were excluded from the citation code. In order to facilitate a better match the cited paper titles were changed to lower case, and any punctuation was removed. An example of the variation in cited paper titles in citations to a single article after this initial formatting is given in Figure 6.6. It can be seen that the remaining errors typically include the omission of plural s (no. 1 and 2), the

---

[56] Resources for running the program as well as the source code of the program written in C++ for Unix was kindly made available to the author for the purposes of constructing the citation code, which is greatly acknowledged.

[57] Also called a Damerau-Levenshtein distance or Levenshtein distance after its inventors (Damerau, 1964; Levenshtein, 1965).

omission of whole words (no. 1 and 3), the inclusion of wrong words (no. 4), as well as various typos and misspellings (no. 6, 7 and 8).

1. stochastic relaxation gibbs distribution and bayesian restoration of images (**)
2. stochastic relaxation gibbs distribution and the bayesian restoration of images (*)
3. stochastic relaxation gibbs distributions and bayesian restoration of images (*)
4. stochastic relaxation gibbs distributions and the bayesian distribution of images (*)
5. stochastic relaxation gibbs distributions and the bayesian restoration of images
6. stochastic relaxation gibbs distributions and the bayesian restoration pf images (*)
7. stochastic relaxation gibbs distributions and the bayesian restorationof images (*)
8. stochastic relaxation gibbs distributions and thebayesian restoration of images (*)

**Figure 6.6. Example of inconsistencies and errors in the cited paper titles (`<atl>`) in citations to Geman and Geman (1984). Capitalisation and punctuation have been removed, and strings with errors are marked with an asterisk (*).**

It was attempted to reduce the effect of these kinds of errors and inconsistencies by the use of the edit distance algorithm in the like program. Edit distance matches were done on the strings in each group, and the results post processed to identify citation to the same documents. Details of this process are given in Appendix 4. Before the corrections there were 85,707 unique strings among the 116,265 cited paper titles with a cited year. After the corrections there were 70,634 unique strings in 7,111 documents (see Table 6.3), corresponding to a reduction of 17.6%. Thereby 15,073 cited paper titles have been replaced with corrected versions. No formal test of the accuracy of the error reduction has been carried out, as no test corpus to compare the faulty and the corrected strings have been available[58]. Of the 70,634 unique citations 17,487 (24.8%) were cited in more than one document. The citation that was mentioned in most documents occurred in 81 different documents.

To ease processing each of the 70,634 unique cited paper titles were assigned a numeric identifier (called unique reference ids - or URIDs). The URIDs for the 7,111 documents in the INEX corpus were formatted in InQuery's pseudo SGML format and subsequently indexed resulting in a citation index. Two citation indexes were generated in InQuery from the URIDs: one conventional index where each citation is listed once for each document it occurs in (called a *flat* citation index), and one where each citation

[58] A small manually compiled test corpus is offered in Lawrence, Giles and Bollacker (1999). The authors were contacted to obtain this corpus, but unfortunately no reply was given.

is listed as many times as it occurs in the full text of the document (called an *expanded* citation index). As discussed in Chapter 1 the creation of the latter is facilitated by the availability of scientific full text documents in electronic format, and is one of the new possibilities offered by this type of documents. The expanded citation index allows investigations of whether or not multiple mentions of a reference in the body text can provide stronger evidence of the value of such references as index and query keys in IR, as investigated on small scale by Herlach (1978). The information for the expanded citation index could be extracted without difficulty from the INEX corpus because the references in the bibliographies are linked to the location(s) in the full text where they are mentioned. Statistics for the flat and the expanded citation indexes are given in Table 6.3. Both indexes cover 7,111 documents with a total of 70,634 unique index keys. The flat citation index has a total of 111,741 index keys, or 9.9 references per article on average. The expanded citation index has a total of 192,881 index keys, or 27.1 references per article on average. Note that because the citation indexes were indexed in InQuery the best match normalisation rules in InQuery's implementation of the tf*idf weighting scheme were also applied to the citation indexes (See Section 6.2.1 and Formula 3 above). The consequence of this is for instance that the weights of the references in the citation indexes were normalised for document length (i.e., the number of references), and that the *idf* value of each reference in the citation index also form part of the weight. Thereby a document will be given a higher weight if it contains a reference that is very rare in the whole index and this reference occurs as seed document in a query because of the *idf* factor. If the document also contains few references compared to the rest of the index it will be given a higher weight in the ranking because of the *tf* factor. The risk that the top ranked documents always consists of, e.g., review articles with a large number of references, is therefore considerably reduced.

### *6.2.5 Database setup*

The representations listed in Table 6.3 were indexed as *separate* InQuery databases. There were two reasons for this choice: Firstly, InQuery does not provide features that can be exploited to run all three steps of the best match boomerang effect within InQuery. For instance, it is not possible to save search sets and combine them with other sets in the same session – thereby the identification of pools and overlaps in Step 2 is not possible. Secondly, it was not possible to modify InQuery to include all representations in separate fields in a single database within the given timeframe. All representations were indexed using InQuery's standard stop words list, as well as

stemming. The stop word list consists of 418 English words, and the stemming is k-stem, an improved version of Porter's stemmer (Porter, 1980). The use of the databases is described in connection with the discussion of each test run in the next section. In general, the use of separate databases for each representation can be expected to perform differently from a single database with all the representations indexed in fields. The reason is that the *idf* values for a given index key will be different for each separate database (See Section 6.2.1 and Formula 3 above). This may be an advantage in application of the theory of polyrepresentation because the term weighting is adapted to the characteristics of each representation. This is not tested in the dissertation, however.

## 6.3 Test runs

The main type of test run in the main experiments is the best match boomerang effect with several variables. In addition two types of baseline runs were constructed for the main experiment. All of the experiments were run using InQuery's batch mode module, and all answer sets consist of the top 100 documents retrieved as in INEX 2002. This threshold is referred to as DCV_step3 to distinguish it from the other thresholds applied in the main experiment. Since the pools in INEX were based on a maximum of 100 documents from each run submission, a value of DCV_step3 larger than this seems unreasonable. If significantly smaller values are used there is a risk of getting results that are unstable and have high error rates (Buckley and Voorhees, 2000). We chose to set DCV_step3 at 100 documents in the main experiments to ensure the largest degree of stability in our evaluation measures. An added benefit of this is that the results can be compared directly with the official results published from INEX 2002. The three types of runs are described below after a description of the queries used in all of them. In addition, details are given in Section 6.3.4 on a number of other runs, including the three runs officially submitted to INEX 2002. These were preliminary runs that were submitted to INEX before the execution of the main experiment. As discussed in Chapter 7 and 8 the experience gained subsequently affected the main experiment – most notably they led to the implementation and testing of further variables in the best match boomerang effect and the polyrepresentation baseline. The variables are described in connection with the general description of the best match boomerang effect and the baselines immediately below in Section 6.3.2 and 6.3.3. The details of the preliminary runs submitted to INEX, and their relation to the main experiment runs, are given at the end in Section 6.3.4.2.

### *6.3.1 Queries*

One of the opportunities offered by the theory of polyrepresentation is to adapt the queries to each of the representations in an attempt to achieve better performance. Indeed one may be forced to do so when the representation of the information need cannot be matched directly against a given representation of the documents as discussed in Section 2.1. It was chosen *not* to adapt the queries to individual representations in the test of the best match boomerang effect, except for the translation of natural language queries into citations codes done by the boomerang effect. Instead the queries were used "as is" for all functional and cognitive representations in the main experiment for a number of reasons.

Firstly, all functional and cognitive representations are textual, and some degree of match can be expected in all representations as all index keys and query keys are stemmed. Secondly, the intention behind the best match boomerang effect is for it to be at the unstructured end of the polyrepresentation continuum, and to rely on the best match principle implemented in InQuery to rank relevant documents at the top of the output. Thus instead of forcing a structure on the queries to achieve a high quality input for the boomerang effect in Step 1, the ranked output of InQuery is used in an attempt to achieve the same. Therefore the queries are not enhanced, e.g., by creating structured or expanded queries as done in Kekäläinen (1999), or by targeting the queries to the individual representations as done in Madsen and Pedersen (2003). Thirdly, no experience, or time to acquire this, was available in using automatic tools to modify the queries.

Three of the four elements in the INEX topics are intended as sources for the generation of ad hoc queries: title, description and keywords population (See the INEX guidelines for topics development (Fuhr et al., 2003, p. 178-181). The longer description element, often in the form of one or two sentences, was judged to be too unrealistic compared to what users might provide to a normal IR system. An IR system with an advanced interactive query model builder might have extracted statements that are similar to the descriptions. However, since the descriptions in the INEX topics have not been extracted in this way, and since no tools for the further processing were available, it was decided not to use the descriptions. As discussed above in 6.1.2 the relatively few words in the title and keyword elements can be regarded as different request versions, and the description element as a kind of problem or goal statement. It was chosen to use the title

and keyword elements for the queries in the main experiment as they resemble the short queries given by users of operational systems. The queries used in all runs in the main experiment therefore consist of a concatenation of the title and keywords elements from the CO topics into a "bag of words". The query keys in the title element are often repeated in the keyword element with the result that some keys will occur several times in the queries. As the default query parser is used, these query keys will be given higher priority in the query network in InQuery, because these keys are automatically weighted by the number of times they occur in a `#wsum` query. The repetition of some of the query keys could have a positive effect on results, although Rajashekar and Croft (1995) were not able to show that it did. Figure 6.7 illustrates CO topics transformed into queries, and InQuery's parsing of the queries on the topics from Figure 6.2 and Appendix 10. Note that stop words like 'of' and 'on' have been removed from the queries, and that the query keys are stemmed automatically by InQuery. The first number of the `#wsum` query is a weight that can be used to give higher weight to all the query keys in the `#wsum` clause – by default it is 1.00 (See Section 6.2.1 above).

CO topics transformed into queries:

```
q39 = Video on demand video, video on demand, architecture, server,
design ;

q36 = Heat dissipation of microcomputer chips heat dissipation
circuit design heat removal heat generation thermal modelling low
power ;

q47 = concurrency control semantic transaction management application
performance benefit "concurrency control" "semantic transaction
management" "application" "performance benefit" "prototype"
"simulation" "analysis" ;

q58 = Location management scheme location management area cell
wireless ;
```

Parsing by InQuery:

```
q39 = #WSUM(1 3 video 2 demand 1 architecture 1 serve 1 design) ;

q36 = #WSUM(1 4 heat 2 dissipate 1 microcomputer 1 chip 1 circuit 1
design 1 removal 1 generate 1 thermal 1 model 1 low 1 power) ;

q47 = #WSUM(1 2 concurrent 2 control 2 semantic 2 transact 2 manage 2
application 2 performance 2 benefit 1 prototype 1 simulate 1
analysis) ;

q58 = #WSUM(1 2 locate 2 manage 1 scheme 1 area 1 cell 1 wireless) ;
```

**Figure 6.7. Example of CO topics transformed into queries, and InQuery's parsing of these. See Figure 6.2 and Appendix 10 for the original topics.**

### *6.3.2 The best match boomerang effect*

The best match boomerang effect was implemented as described in Section 5.4. All eight functional and the two cognitive representations listed in Table 6.3 were indexed in separate representation databases in InQuery as described in Section 6.2.5. The topics were transformed into queries as described above and were run in batch mode against each representation database. In the main experiment the '`norankall`' option was set in the batch mode module to prevent InQuery from ranking the whole database. As a true best match system InQuery's batch mode module ranks all documents in a database in relation to a query by default. This means that the top of the ranked output will contain documents with one or more query keys as expected, but also that documents

without the query keys can appear on the bottom of the rank. If the '`norankall`' option was not set, documents that are unrelated to the query keys might therefore be included in the set of retrieved documents if there are fewer documents with the query keys than the document cut-off value (DCV). This is a particular concern in the best match boomerang effect as the documents retrieved in Step 1 determine which references will form the citation pools in Step 2. A threshold, referred to as the DCV_step1 threshold, was implemented so that the size of the sets of retrieved documents in Step 1 could be controlled as a variable in the experiments. Although the best match boomerang effect is at the unstructured end of the polyrepresentation continuum and do not rely on, e.g., structured queries, the DCV_step1 threshold makes it possible to use only the top ranked documents from InQuery in Step 1. Nine values of the DCV_step1 threshold were tested in the main experiment: 2, 4, 8, 16, 32, 64, 128, 256, and 512. The output of the batch mode module is a list of documents, ranked descending by their belief values as calculated by InQuery, and saved to a file.

A special program, running outside InQuery, was written to takes these files as input and extracts the references from all the documents using either the flat and expanded citation index described in Section 6.2.4. From this the program generates the citation pools in Step 2 for each of the ten representations used in Step 1 (See Figure 5.6). The program also identifies the frequency of occurrence of each citation in the pools, their occurrence across pools (i.e., it identifies the ones in the overlaps between pools), and calculates the final weights for the citations as illustrated in Figure 5.7. If the extended citation index is used, the number of times a reference is mentioned in the body text forms part of the calculation. The final result of the program is a single list of citations, weighted by their frequency of occurrence in the overlaps in Step 2 as described in Section 5.4[59]. The list is ranked decreasingly after the weight of the citations. A threshold, referred to as CCV_step2 (citation cut-off value), was implemented so that the size of the weighted list of citations outputted as result of Step 2 could be controlled as a variable in the experiments. Some of the uncertainty introduced by the automatic

[59] We were warned that InQuery's batch mode module might disregard anything after a decimal point, e.g., parse #WSUM(1 1.75313 URID046918 1.01264 URID035182) as #WSUM(1 1 URID046918 1 URID035182). We therefore multiplied the weights with 1000. This later turned out not to be a problem in InQuery version 3.2, and the multiplication did not affect the results.

translation of the information needs into weighted seed documents might be reduced because only the strongest evidence would be used when the threshold is used. Three levels of the CCV_step2 threshold were tested in the main experiment: Low, Medium, and High, corresponding to top 25 %, 50 %, and 75 % of the total number of citations resulting from Step 2. If, for instance, the result of Step 2 is 13,000 weighted citations ranked in decreasing order, use of the Low threshold will reduce the output to the top weighted 3,250 citations.

The weighted list of citations was formatted as a weighted sum query and submitted to InQuery's batch mode module against either the flat or expanded citation index. Appendix 5 shows an example of such a query. The number of query keys in these lists tended to be quite large with more than 10,000 query keys in some cases. The final result of the best match boomerang effect is the top 100 documents returned for each query in Step 3.

The variables tested in the experiments with the best match boomerang effect were:

1. The *document cut-off value* in Step 1 (DCV_step1),
2. The use of either a *flat* (f) or an *expanded* (x) *citation index* to extract citations for Step 2,
3. The *citation cut-off value* in Step 2 (CCV_step2), and
4. The use of either a *flat* (f) or an *expanded* (x) *citation index* to run the weighted citation queries against in Step 3.

### *6.3.3 Baselines*

As discussed in Section 6.1.4 it was initially a concern whether the recall base in INEX would be adequate since too few whole documents might be assessed. Therefore, in order to ensure an adequate baseline, it was decided to submit two baseline runs in addition to a best match boomerang effect run as part of the TAPIR project's participation in INEX. As interesting results were obtained with the baselines in INEX 2002 they were also included in the main experiment in a slightly modified form. The baselines as used in the main experiment are described immediately below. Details of the parameters used with the baselines in the preliminary INEX submission are given in Section 6.3.4.2 below.

#### 6.3.3.1 The polyrepresentation baseline

The idea behind the polyrepresentation baseline was to examine the effect on the results when employing the theory of polyrepresentation on its own, without exploiting the network of references and citations as in the best match boomerang effect. The polyrepresentation baseline allows the posing of questions such as "Are the results obtained by the best match boomerang effect mainly due to the exploitation of the citation network, or to the theory of polyrepresentation?" The inclusion of the polyrepresentation baseline in the main experiment also allows analyses of the differences between the documents retrieved by it and other approaches, e.g., to study if the same documents are retrieved by the polyrepresentation baseline and the best match boomerang effect. In order to be comparable to the implementation of the best match boomerang effect as described in Section 6.3.2 the polyrepresentation baseline was at the unstructured end of the polyrepresentation continuum.

The polyrepresentation baseline corresponds to Step 1 in the best match boomerang effect, where the same queries are run against the ten different representation databases with the InQuery batch mode module. However, instead of exploiting the resulting answer sets to generate citation pools, the answer sets were *fused* to create a simple polyrepresentation baseline. This was achieved by a program that executed the following steps: First, each of the 10 output files from the batch mode module were parsed, and the ID number and belief value for each document as calculated by InQuery were saved in a data structure. Then, if a document occurred in more than one output file the belief values were added into a final weight for that document. If the document only occurred in one output file the final weight of the document was the belief value in that file. As the last step, all documents were ranked according to the final weights, and the top 100 documents returned as the result of the polyrepresentation baseline. A threshold was implemented so that the size of the 10 sets of retrieved documents could be controlled as a variable in the experiments. This threshold is the same as the DCV_step1 in the best match boomerang effect.

In the polyrepresentation baseline documents that occur in the answer sets of several representation databases, i.e., was in the overlap between them, were given higher priority, as the belief values from all these representations were added into the final weight. Thereby an effect, that is similar to the one in Step 2 of the best match boomerang effect, was achieved: the top ranked documents could consist of, e.g.,

documents with medium belief values from all or most answer sets, or alternatively docs with large belief values from a few answer sets. The approach taken in constructing the polyrepresentation baseline resembles that used in data fusion in IR (See, e.g., Belkin et al., 1995)

In order to be able to make direct comparisons, the same queries were used as in the best match boomerang effect and the bag-of-words baseline. One variable was tested in the polyrepresentation baseline: The DCV_step1 that controls the size of the answer sets from which belief values are fused together.

#### 6.3.3.2 The bag-of-words baseline

The idea behind the bag-of-words baseline was to mimic the traditional "bag of words" approach to best match IR where the document representation consists of one index with all representations as one "bag of words". The bag-of-words baseline allows the posing of questions such as "What is the performance of the best match boomerang effect compared to what can be achieved by traditional methods?". The inclusion of the bag-of-words baseline in the main experiment also allows analyses of the differences between the documents retrieved by it and other approaches, e.g., to study if the same documents are retrieved by the bag-of-words baseline and the best match boomerang effect.

An additional index was constructed for the bag-of-words baseline which contained all of the functional and cognitive representations in Table 6.3 in the same field. Statistics of this mixed representation are given in Table 6.4.

**Table 6.4. The bag-of-words index. The number of documents, number of unique index keys, and the total number of index keys are derived from InQuery, i.e., they do not include stop words and have been stemmed (See Section 6.2.5). The size is the number of MB of the representation before indexing, but excluding tags and ID numbers.**

| Type | Name | No. of documents | No. of unique index keys | Total no. of index keys | Size in MB |
|---|---|---|---|---|---|
| mixed | bag-of-words | 12,106 | 54,426 | 4,415,541 | 47.2 |

In order to make a direct comparison between the two the queries used in the bag-of-words baseline are the same as those run against the best match boomerang effect (as described in Section 6.3.1). The final output of the bag-of-words baseline is the top 100

documents returned by InQuery for each query. No further variables are tested in the bag-of-words baseline.

### *6.3.4 Other runs*

#### 6.3.4.1 Individual representations

In order to learn more about the effect on the results of each of the functional and cognitive representations listed in Table 6.3 a run was made with *each* of the 10 representations. The same queries as used in the other runs were also used and no further variables were tested in these runs. Because the best match boomerang effect as implemented in the main experiment is at the unstructured end of the polyrepresentation continuum individual overlaps between the representations are not isolated. Therefore the effect of strictly identified overlaps cannot be studied as done in Madsen and Pedersen (Madsen and Pedersen, 2003). Although individual overlaps between the representations are not studied explicitly, as they would be in applications based on the structured end of the polyrepresentation continuum, the relative contribution of each representation in Table 6.3 is interesting as it might provide indications as to which of the representations that have the potential to improve retrieval when used in a polyrepresentative manner. Alternatively, if one or more representations show inferior performance it might be considered to merge such representations with other representations at the indexing stage, as done in Madsen and Pedersen (2003).

#### 6.3.4.2 Official INEX 2003 runs

Three runs were submitted to INEX 2002: a best match boomerang effect run, a polyrepresentation baseline run, and a bag-of-words baseline run. Apart from using the description element from the INEX CO topics in addition to the title and keywords for the queries, the three runs were implemented as described above in Section 6.3.2 and 6.3.3, but only with a specific combination of the variables. The following combinations of variables were used: both the polyrepresentation baseline and the boomerang run were run with a fixed DCV_step1 of 500 documents. The boomerang run used all citations from Step 2 in the forward chaining, that is, no CCV_step2 threshold was invoked, and the expanded citation indexes were used in both Step 2 and Step 3.

## 6.4 Performance evaluation

A wealth of methods to measure the effectiveness of IR techniques have been proposed in the literature. The prevalent performance measures are based on *precision* and *recall*.

Precision is defined as:

$$precision = \frac{number\ of\ relevant\ documents\ retrieved}{total\ number\ of\ documents\ retrieved} \quad (5)$$

Recall is defined as:

$$recall = \frac{number\ of\ relevant\ documents\ retrieved}{total\ number\ of\ relevant\ documents\ in\ the\ collection} \quad (6)$$

(Cleverdon and Keen, 1966, p. 31). Retrieval results are commonly presented as either 1) average precision vs. recall (P-R) curves, or 2) as average precision and recall at a fixed DCV (Hull, 1993, p. 330). In the latter, a fixed number of documents are retrieved for each query and precision and recall values are calculated and averaged over requests. In the P-R curves precision is measured at fixed standard recall values, e.g., at ten recall points (0.1, 0.2, …, 1.0 recall). Because there is often no exact precision value for the standard recall levels in relation to a given query, the P-R curves are averaged over queries by interpolation (van Rijsbergen, 1979, p. 150-153).

Since precision and recall describe the IR performance over a number of values they constitute a bivariate measure of retrieval effectiveness. In order to compare two retrieval techniques one of the variables is fixed (as the independent variable) and performance is judged in relation to the other (as the dependent variable). In the P-R curves recall is the independent variable and precision is the dependent variable. The assumptions behind the P-R curve are that when precision is averaged across queries a particular recall level must be achieved by every query, and that the best technique is the one that does this with the fewest number of non-relevant documents on average. P-R curves may therefore be said to measure the *equivalence between queries*. Note that the curve does not provide any information as to the *number* of documents that must be retrieved in order to achieve a given level of recall, and that 100% recall may not be

achieved in every query. Because the number of relevant documents are different for individual queries a recall of 50% may result in, e.g., 10 documents for one query and 100 documents for another query. Based on the P-R curves different IR techniques can be compared visually on the curve, or on the basis of their average precision at a given level of recall, e.g., the precision at .5 recall. (Hull, 1993, p. 331).

In the DCV-based method the fixed number of documents is the independent variable, and precision and recall are the dependent variables. From a user's viewpoint this can be seen to represent the effort that she is willing to make by examining a number of documents for each query, and the DCV-based method may be said to be based on *equivalent effort* rather than equivalence between queries. Because of the fixed DCV, precision cannot reach 100% if the number of relevant documents is smaller than the DCV, which may lead to misleading results when comparisons between queries and IR techniques are based on average precision values. Similarly, 100% recall cannot be reached if the number of relevant documents is larger than the DCV, leading to problems with the interpretation of average recall values. Hull therefore suggests to measure precision and recall over a range of DCVs, and then average the results, which will smooth out any irregular effects caused by using a particular DCV (Hull, 1993, p. 330-332).

An important question in selecting a performance measure is whether the assumptions of the measure are related sufficiently to the *intended function* of the tested IR technique. In measures based on P-R curves recall is central, but recall may not be important for most users of ad hoc IR systems. Hull (1993) believes that averaging on the basis of equivalent effort, i.e., using the DCV-based method over a range of DCVs, is much more reasonable for most IR applications. Seen from a cognitive viewpoint this is also the perspective to evaluate: the theory of polyrepresentation is intended to provide high precision by pushing relevant documents to the top of the ranking for the user to inspect (See, e.g., the Appendix in Ingwersen (1996)). Another important question is the stability and reliability of the chosen measure. Buckley and Voorhees (2000) have tested the error rate and discrimination power of five of the commonly used performance measures, including DCV-based ones and a P-R based one (precision at .5 recall). Neither of these had low error rates compared to the others, and the DCV-based ones had some of the highest error rates especially for low DCVs, as could be expected. Buckley and Voorhees did not, however, average the results over several DCVs in their

test as recommended by Hull (1993) and therefore we do not know the error rates associated with this approach. The best measures for general purpose retrieval were *average precision* (calculated as the mean of the un-interpolated precision scores obtained after each relevant document has been retrieved), and *precision after R documents* (where R is the total number of relevant documents for the current topic). Both of them have low error rates and were not noticeably affected by a small topic set size, but precision after R documents did not have as much discrimination power (measured as the number of ties between IR techniques) as average precision.

The decision by the INEX organisers to use non-binary relevance assessments allows calculation of some of the novel performance measures that have been proposed in recent years. As discussed above, the evaluation of IR systems has primarily been based on measures of recall and precision. By definition (See Formula 5 and 6) these two measures require a dichotomous partition of the relevance assessments into those documents that are relevant and those that are not. Therefore, until recently, even when graded relevance assessments were collected these have typically been collapsed into a binary scale to facilitate calculation of recall and precision (See, e.g., Blair and Maron, 1985; Saracevic et al., 1988; Su, 1992; Pao, 1993). This situation has changed in recent years with the proposal and application of new performance measures that can exploit non-binary relevance assessments.

As part of her work on establishing an evaluation methodology for interactive IR systems, Borlund identifies a need for alternative performance measures "…capable of bridging the interpretative distance between the objective and subjective types of relevance…as well as managing non-binary assessments…" (Borlund, 2000a, p. 153). As an example of such measures she proposes two performance measures that can make use of non-binary relevance assessments: the relative relevance measure (RR) and the ranked half life (RHL) indicator (Borlund and Ingwersen, 1998; Borlund, 2000a). The purpose of the RR measure is to describe the degree of agreement between different classes and types of relevance when these are available in interactive evaluations of IR systems. The RHL indicator compares the performance of IR systems based on their ability to push highly relevant documents to the top of the ranking. As just two types of relevance are available in the non-interactive laboratory experiments with the INEX test collection, only the RHL indicator will be discussed here. RHL is proposed as a supplement to the traditional measures based on recall and precision, and is related to

the expected search length (ESL) measure proposed by Cooper (1968) because it takes the position of relevant documents in the ranking into consideration. Based on an accumulation of the relevance assessments the RHL indicator calculates the median point of these values at a given document cut-off value (DCV). Different degrees of relevance can be weighted in the accumulation, e.g., by assigning a value of 1.0 to highly relevant documents, and a value of 0.5 to maybe relevant documents. If highly relevant documents tend to be at the top of the ranked list, the RHL value will be higher than when highly relevant documents are scattered evenly or are at the lower half of the ranking. The RHL indicator complements simple calculations of precision by indicating how well an IR technique pushes the highly relevant documents to the top of the rank at a given level of precision, and in that sense "…supplies additional information about the degree to which the engine is capable of ranking its output according to user-oriented relevance." (Borlund, 2000a, p. 153).

In a series of papers Järvelin and Kekäläinen propose a number of evaluation methods that can exploit non-binary relevance assessments, and demonstrate their use in a number of case studies (Järvelin and Kekäläinen, 2000; 2002; Kekäläinen and Järvelin, 2002b). Three types of methods are proposed:

1. Calculation of separate recall bases for each degree of relevance, resulting in
    a. a novel application of P-R curves with separate curves for each degree of relevance, and
    b. average precision calculations at each degree of relevance.
2. Calculation of generalized precision (gP) and generalized recall (gR) values based on non-binary degrees of relevance.
3. Calculation of a new performance measure which computes the cumulative gain a user achieves by inspecting the ranked retrieval output up to a certain point,
    a. either directly as cumulated gain (CG), or
    b. as discounted cumulative gain (DCG), where the relevance values are progressively reduced as the rank increases as a modelling of user persistence, or
    c. as normalised CG (nCG) or DCG (nDCG) relative to the theoretically best obtainable performance with the recall base in a given test collection.

The main argument for proposing the methods is that "…evaluation methods should credit IR methods for their ability to retrieve highly relevant documents…" because the

number of relevant documents may easily exceed the number of documents a user is willing to investigate in large, modern IR systems (Järvelin and Kekäläinen, 2000, p. 41). The risk is that "…differences between sloppy and excellent retrieval methods may not be observed in evaluation…" when binary relevance assessments are used, because these may include documents that are only marginally relevant in the documents judged as relevant. Sormunen (2002) has investigated if this is the case by reassessing 38 topics from TREC-7 and TREC-8 using a relevance scale with four degrees of relevance. He found that more than half of the documents judged as relevant in TREC were regarded as marginally relevant in the reassessment. Thus the use of evaluation methods that distinguish between different degrees of relevance like those proposed by Borlund and Järvelin and Kekäläinen is supported for evaluations of IR techniques where the retrieval of highly relevant documents rather than marginally relevant is important.

The first two types of methods extend existing evaluation methods by making it possible to incorporate non-binary relevance assessments. Differences between IR techniques can be studied separately at each degrees of relevance with the new P-R curves (type 1.a) and the related average precision calculations (type 1.b). Generalized precision and recall figures (type 2) can be used in construction of the traditional performance measures as described above, e.g., for calculating averages over queries, precision over recall levels or at various DCVs and the drawing of performance curves. The graded relevance scores are normalised to be in the interval between 0 and 1, e.g., by dividing all scores by the highest score. Both ordinal and continuous scales are supported. Alternatively a non-linear set of weights can be specified, for example to place more emphasis on highly relevant documents.

Although related to several measures proposed earlier (see below), the measures based on the idea of cumulated gain (3.a-c) represent a novel contribution. The (D)CG measures are explicitly user-oriented in that they calculate the gain cumulated by inspecting an explicit number of documents moving down from the top of the rank. In addition they can emphasise documents with high degrees of relevance as well as documents that are retrieved early in the rank. Cumulated gain, CG (type 3.a), is calculated by substituting document IDs by their relevance values and accumulating these beginning from the top of the rank down to a specific point. If for example the relevance scale with four degrees of topical relevance from INEX is used (see Section 6.1.3 above), and five documents have been retrieved these might be represented by

their relevance scores as a vector: G' = (3, 2, 0, 1, 2). This means that the first document has been assessed as Highly relevant, the second and fifth documents as Fairly relevant, the fourth as marginally relevant, and the third as irrelevant. The original relevance scores may be used, or a different set of weights or *gain values* may be used. The cumulated gain at certain rank position *i* is calculated by summing from position 1 to *i* in the vector. If rank position *i* in the gain vector G is denoted formally by G[*i*], CG can be defined recursively as the vector CG where (Järvelin and Kekäläinen, 2002, p. 425):

$$CG[i] = \begin{cases} G[1], & \text{if } i = 1 \\ CG[i-1] + G[i], & \text{otherwise} \end{cases} \quad (7)$$

Thus from G' we obtain CG' = (3, 5, 5, 6, 8) with the relevance scores from INEX used directly as gain values. The cumulated gain can be read at any given rank directly, e.g., CG is 7 at rank 4. In this way the cumulated gain measures may be said to be user-oriented as the actual gain obtained by inspecting an explicit number of documents retrieved by a particular IR technique is immediately obvious. Direct comparisons between two or more IR techniques are also possible by comparing the vertical distance at any rank – the distance represents the wasted effort by not using the best technique. In addition, an *ideal vector* can always be calculated based on the recall base (see below). The cumulated gain of the ideal vector can then be used as a baseline against which to compare different IR techniques. If, for instance, the recall base contains 3 Highly relevant documents and 2 Fairly relevant documents, the ideal vector would be I' = (3, 3, 3, 2, 2), and the cumulated gain of the ideal vector $CG_I'$ = (3, 6, 9, 11, 13).

Discounted cumulated gain, DCG (type 3.b), is calculated by progressively reducing the CG vector by a rank-based discounting factor to model user persistence. Thereby relevant documents retrieved late in the rank will not count as much as relevant documents retrieved early in the rank. In the DCG the document score is divided by the log of its rank in order to provide a smooth function that is not too steep for progressively reducing the influence of documents retrieved at higher rank positions.

Division by, e.g., the rank itself would be too severe. If *b* denotes the base of the logarithm the discounted cumulated gain vector is defined recursively as the vector DCG where (Järvelin and Kekäläinen, 2002, p. 425):

$$DCG[i] = \begin{cases} CG[1], & if\ i < b \\ DCG[i-1] + G[i] / {}^{b}\log i, & if\ i \geq b \end{cases} \quad (8)$$

The discounting factor *b* may be varied to simulate different levels of user persistence. For example, an impatient user who does not wish to examine documents a long way down the rank could be modelled by $b = 2$, whereas a persistent user could be represented by $b = 10$. With $b = 2$ and the same data as in the example above we get DCG' = (3, 5, 5, 5.50, 6.43).

To calculate the nCG and nDCG measures (type 3.c) it is necessary to construct the theoretically best possible (ideal) vector as a baseline. Instead of unrealistically assuming that this vector consists of only highly relevant documents the recall base is used as basis for the ideal vector. It is constructed by first listing all the Highly relevant documents, followed by the Fairly relevant, and finally the marginally relevant. Irrelevant documents are represented by zeros at the end of the ideal vector. The ideal vectors can be used to construct a baseline for the CG and DCG measures by dividing the (D)CG value at position *i* with the value of the ideal vector at position *i*. The resulting values are the normalised cumulated gain (nCG), and the normalised discounted cumulated gain (nDCG). Note that the CG and DCG values can be compared directly to the ideal vector as an alternative to calculating nCG and nDCG measures. The ideal vector can also be utilised to measure the quality of a particular IR technique: The area between the normalised ideal vector and the normalised vector for the technique represents the quality of the technique. The difference between the normalised vectors of two IR techniques may also be compared: "The average of a (D)CG vector (or its normalized variation), up to a given ranked position, summarizes the vector (or performance) and is analogous to the un-interpolated average precision of a DCV curve up to the same ranked position." (Järvelin and Kekäläinen, 2002, p. 427).

Using the three types of cumulated gain measures Järvelin and Kekäläinen (2000; 2002) demonstrate the calculation of averages over topics, the drawing of graphs, how to make comparisons between two or more IR techniques, as well as procedures for testing the

statistical significance of the results. The cumulated gain measures are related to a number of the measures proposed in the literature, e.g., Cooper's expected search length (1968), Pollack's sliding ratio (1968), as well as Borlund's RHL discussed above (1998). Järvelin and Kekäläinen argue that the cumulated gain measures have a number of advantages over previously proposed measures (2002, p. 429), including the fact that they combine the degree of relevance and their rank in a coherent way, that they give an estimate of the cumulated gain as a single measure at a given rank regardless of the size of the recall base, that they are not heavily dependent on outliers, and that they are obvious to interpret. In addition, the discounted versions realistically weight down the gain for documents at higher ranks, and allow for modelling of user persistence. Finally, the normalised versions present the cumulated gain of different IR techniques relative to the actual recall base including graded assessments, and thereby facilitate analysis of performance differences. The main disadvantages of the cumulated gain measures are that there is no general justification for setting the parameter values of the measures. The parameters include the last rank to be considered (i.e., the DCV), what gain values to use, and the actual discount factor to employ. While asserting that the evaluation context and scenario should suggest these values in practice, Järvelin and Kekäläinen recognise that especially the choices of gain values and discount factor are somewhat arbitrary (Järvelin and Kekäläinen, 2002). Note that by calculating separate recall bases for each degree of relevance the problem of selecting gain values is avoided in the novel P-R curves and related average precision calculations (type 1.a and 1.b above). These measures thereby remain true to the statistical restrictions set by ordinal relevance scales. In contrast, the generalized precision as well as the cumulated gain measures (type 2 and 3) break the ordinal scale by allowing inferences like "…a document of relevance degree 3 is three times as relevant as a document of relevance degree 1." (Kekäläinen and Järvelin, 2002b, p. 1122). This is a conscious choice by the authors, because it allows the calculation of single, combined measures that can be used to pose "what if … ?" questions, like what happens when, e.g., highly relevant documents are emphasised while lower relevance degrees still are allowed to influence the measure of performance to a certain degree.

A potential problem with basing the evaluation on highly relevant documents is that there might be too few of them. IR performance measures are inherently unstable when there is only a few relevant documents for a topic because a small change in document ranking can produce a large change in the performance score (Voorhees, 2000). This

may result in a situation where it is difficult to conclude whether two IR techniques are significantly different from one another (Buckley and Voorhees, 2000). Voorhees (2001) has validated the use of the cumulated gain measure using data from the TREC-9 web track with this problem in mind. She shows that the relative performance of IR techniques "…change when evaluated by highly relevant documents as compared to when evaluated by generally relevant documents." (2001, p. 81). In her study she varies the weight given to the highly relevant documents from 1 to 1000 in calculations of DCG, but keeps the discounting factor fixed. She concludes that there may be a risk that DCG measure with very large gain values will be unstable when only few highly relevant documents are available: "Using DCG with smaller [gain] ratios (say, 3-5) will incorporate all relevance information into the score to increase stability, while still rewarding systems for retrieving highly relevant documents first." (Voorhees, 2001, p. 78).

### *6.4.1 Performance measures in the main experiment*

Several performance measures were considered for the analysis of the main experiment. It was decided to make use of the graded relevance assessments rather than conflating them into binary form. This will provide a broader perspective on the results by making it clear whether there is noticeable variation at different degrees of relevance, with the result that hopefully more can be learned from the experiment.

Two main types of performance measures were chosen to study two different aspects of performance in the main experiment:

1. Traditional P-R curves over 100 recall points based on generalized precision (gP) and generalized recall (gR), with average precision calculated over these 100 recall points,
2. DCG and nDCG curves as well as averages derived from these.

The first type of measure studies performance in terms of the tested IR technique's ability to reach high recall with the fewest number of irrelevant documents. Although average precision over a range of DCVs can be regarded as better related to the users' behaviour (Hull, 1993) than the P-R curves, the latter is preferred because they seem to be more stable and reliable (Buckley and Voorhees, 2000). Referring to the results of an earlier study Voorhees states that "The robustness of retrieval evaluation across different sets of human assessors' judgements is due mostly to the effect of averaging

over many topics. [(Voorhees, 2000)]" (Voorhees, 2001, p. 77). This robustness of the chosen performance measures is particular important with the CO topics in the INEX test collection as relevance assessments are available only for 24 of the CO topics. The second type of measure based on cumulated gain is more user-oriented in that the cumulated gain measures are much closer related to the effort a user is willing to make by examining documents starting from the top of the rank as discussed above.

The INEX organisers supplied an evaluation tool, `inex_eval`, which provides two perspectives on the results: curves that plot recall vs. precision (either as P-R curves, or as recall and precision at fixed DCVs), and average precision values based on the same data. The calculations of all measures in `inex_eval` are based on gP and gR, as defined by Kekäläinen and Järvelin (2002b), and are calculated over 100 recall points. The average precision values are calculated over these 100 recall points after each retrieved document. The average precision measure tested by Buckley and Voorhees (2000) as described above is similar to the average precision measure computed by `inex_eval`. The latter will often result in the same relative differences between IR techniques although the absolute values will be lower. The average precision measure calculated by inex_eval is, however, interpolated – a theoretical justification for this is given in Gövert and Kazai (2003, p. 12). The combined effect of the generalized precision and recall figures as basis for the average precision calculations is somewhat like that proposed in the RHL indicator and the cumulated gain measures: IR techniques that retrieve highly relevant documents early in the rank are rewarded. Two default quantification functions are implemented in `inex_eval` for the calculation of gP and gR: a *strict* and a *generalized* function (See Figure 6.8). The strict function (which we call str_inex) only considers documents assessed with 3E as relevant – in effect this collapses the assessments into a binary distribution similar to the ones used in traditional test collections[60]. The generalized function (which we call gen_inex) exploits the two relevance dimensions in INEX to assign higher weights to documents with higher degrees of relevance, and better coverage.

[60] Although the assessments in TREC are probably better described by a function where documents that are highly, fairly, and marginally relevant are considered as relevant (Sormunen, 2002).

| Quantification function | INEX default relevance scores |
|---|---|
| strict (str_inex) | 1.00 if 3E |
| | 0.00 if anything else |
| generalized (gen_inex) | 1.00 if 3E |
| | 0.75 if 2E, 3L |
| | 0.50 if 1E, 2L, 2S |
| | 0.25 if 1S, 1L |
| | 0.00 if 0N |

**Figure 6.8. The default relevance scores of the str_inex and gen_inex quantification functions used to calculate generalized recall and precision values in the official INEX2002 results (Gövert and Kazai, 2003, p. 11).**

Since the main interest of almost all INEX participants is the retrieval of components rather than whole documents the default quantifications in Figure 6.8 reflect this. This is unfortunate in relation to the IR techniques tested in the main experiment because only whole documents can be retrieved with these. The relevance propagation rules discussed in Section 6.1.3 have two consequences for IR techniques that retrieve whole documents: Firstly, the highest assessment of the topical relevance dimension of any component in the documents is propagated upwards to the level of the whole document. Secondly, if any component other than the whole document is assessed as having exact coverage, the whole documents will be assessed as being too large. The first consequence has no negative effects on IR techniques that retrieve whole documents, but the second does. For instance, if the entire body text of a document has been assessed as 3E, the whole document will automatically be assigned the score 3L. Therefore, when using the str_inex function even documents with large portions of relevant material will be ignored. A more reasonable function in relation to IR techniques that retrieve whole documents would be one that does not assign lower weights to whole documents that are considered too large *and* relevant (of any degree). Figure 6.9 shows two such functions that are analogous to the ones defined by INEX, but adapted for whole documents.

| Quantification function | Relevance scores for IR techniques that retrieve whole documents |
|---|---|
| strict (str_whole) | 1.00 if 3E, 3L<br>0.00 if anything else |
| generalized (gen_whole) | 1.00 if 3E, 3L<br>0.66 if 2E, 2L<br>0.33 if 1E, 1L<br>0.00 if 2S, 1S, 0N |

**Figure 6.9. Relevance scores of the str_whole and gen_whole quantification functions modified to calculate generalized recall and precision values that are more reasonable in relation to IR techniques that retrieve whole documents.**

The strict function (str_whole) considers any document as relevant if it is assessed as highly relevant (i.e., 3E or 3L). This is justified by the fact that no 3S assessments are allowed in INEX, and therefore even if the topic of interest is only a minor theme in the document it will still have been discussed exhaustively (See the definitions of the two relevance dimensions in Figure 6.3 and Figure 6.4 above). The generalized function (gen_whole) in Figure 6.9 considers documents that are too small as irrelevant, and puts equal weight on documents that have exact coverage or are too large. In effect, the str_whole and gen_whole functions largely ignore the coverage dimension, except for the documents that are assessed as being too small. The result of the str_whole and gen_whole functions is an optimistic relevance propagation that, in relation to the degree of relevance, rewards whole documents of which not a too diminutive portion has been assessed as relevant. Relevance propagation implemented in this way seems justified by the evaluation scenario of the main experiment.

The quantification functions in both Figure 6.8 and Figure 6.9 will be used to generate a total of four sets of P-R curves over 100 recall points and related average precision calculations. The official INEX quantification functions will be used in order to maintain a direct comparability with the published result from INEX2002, and the modified functions in Figure 6.9 will be used as they are better suited for the type of IR systems tested in the main experiment. Average precision figures will be calculated for all runs and P-R curves presented for the most interesting ones as these are readily available from `inex_eval`. nDCG and DCG curves and figures will be presented for the most interesting runs only, as no program was available to calculate the data for all

runs in a convenient manner. The topical relevance assessments are used directly as gain values for the calculation of the nDCG and DCG measures and the coverage dimension is ignored. The natural logarithm (based on the constant e = 2.718) is used as discount factor as also done by Voorhees (2001). This is a rather strict discounting factor, but it is preferred as a realistic modelling of user persistence. The last rank to be considered is of some importance in the cumulated gain measures as there may be differences between two systems at low and high ranks, and this will affect average gain values. Järvelin and Kekäläinen recommend to produce (D)CG curves with quite high DCVs, and to test for differences between IR techniques at lower DCVs if the techniques appear to perform differently (Järvelin and Kekäläinen, 2002, p. 438). We will draw the (n)DCG curves and inspect them to see if there are differences to be noted, and test for them if present. We will not examine (n)DCG past rank 100 since the recall base has been built from submissions with a maximum DCV of 100. In addition, differences among the individual queries are illustrated by histograms where the average precision of each query is plotted against the median average precision of all INEX 2002 CO topics as a baseline. The latter data has been made available by Norbert Gövert, and will only be shown for the official INEX quantification functions as the median data only covers these functions.

## 6.5 Statistical testing

In laboratory experiments one will usually want to know if empirically identified differences between IR techniques are statistically significant. If some techniques can be shown to be significantly better than others then the former would be good candidates for implementation or further research. However, the characteristics of the measures used to evaluate IR techniques, such as those described above, are not ideally suited for statistical testing. Robertson (1981) points out that sampling is problematic in relation to both documents and information needs as none of them can be regarded as random samples in a formal sense. This applies especially for information needs as it is usually much easier to get hold of large quantities of documents than information needs. In INEX it is quite clear that the topics are artificial and that they have been carefully selected rather than drawn randomly from a population (See, e.g., the INEX guidelines for topics development (Fuhr et al., 2003, p. 178-181), and Gövert and Kazai (2003, p. 6-7) ). One way of attempting to ensure that a set of artificial topics is representative of a real situation is to try controlling some of the variables associated with queries, but

this is hampered by the fact that we do not know which characteristics of real queries to reproduce (Robertson, 1981, p. 28). As a minimum, the characteristics of the topics and their creation should be described so that they can be compared to what is known about information needs in general (Kekäläinen, 1999, p. 80). Since there might be uncontrollable biases in any one test collection, a better way of ensuring reliable results is to test on several test collections if possible (Robertson, 1981; Hull, 1993; Buckley and Voorhees, 2000). Keen draws attention to the fact that it is important to distinguish between *statistical significance* and *practical significance* (Keen, 1992). Sparck Jones (Sparck Jones, 1974) proposes a rule of thumb, where a difference in mean performance would be regarded as material if it is larger than 10 %, noticeable if between 5 %, and 10 %, and not noticeable if less than 5 %. Buckley and Voorhees have recently validated this rule of thumb for TREC data (Buckley and Voorhees, 2000).

The more powerful statistical inference tests are *parametric*, that is, they are based on assumptions about the population from which the sample is drawn (for instance, that the underlying population follows a normal distribution). Unfortunately, many of the variables measured in IR experiments (e.g., recall) do not satisfy these criteria, and less powerful non-parametric tests have to be used instead (Robertson, 1981, p. 25). The use of non-parametric tests in IR experiments is also recommended by van Rijsbergen (1979) and Keen (1992). Hull discusses the assumptions of parametric tests in relation to IR experiments, and argues that with sufficient data "…there is no reason why researchers cannot test their data against the common distributional assumptions." (1993, p. 333). He recommends to examine the suitability of any particular set of test data for parametric tests, e.g., by diagnostic plots or by the number of ties in the data. If the results of this are positive he recommends using the parametric tests as they have more power.

Non-parametric tests were used in the analysis of the main experiment because the data did not seem to fulfil the requirements for parametric tests. For instance, the number of ties between the techniques were large, indicating that the errors were not continuous. Non-parametric tests are available for comparisons either between two methods, or between more than two methods. For two techniques *the sign test* is a suitable non-parametric test. If more than two techniques are compared, as in the main experiment, one should not simply make pair wise test between all of them. The reason is that the more paired tests that are conducted at the same significance level, the greater the risk

that one of the paired tests will produce a significant result when there is no actual difference between the techniques (Hull, 1993, p. 334). Instead, it is better to use a test that is designed to examine several techniques simultaneously, for instance *the Friedman test* for non-parametric data.

The Friedman two-way analysis of variance by ranks examines differences between the different treatments (in this case IR techniques) based on the relative performance of the treatments within each query, and corrects for the effect of the queries at the same time. If only two treatments are tested the Friedman test reduces to a sign test. The Friedman test requires that the data is at least on an ordinal scale (that is, the test is based on ranked data). The assumptions of the test is that the IR techniques as well as the queries are independent, i.e., that the IR techniques do not influence each other, and that the queries are not obviously related (although slight violations of the latter is not generally a problem according to Hull (1993)). These conditions seem to be fulfilled in the main experiment, and since differences between more than two IR techniques were tested in the main experiment, the Friedman test was used to test the significance of the differences. It tests whether *k* paired samples or treatments have been drawn from the same population or populations with the same mean. The data are cast in a two-way table with *N* rows (representing topics) and *k* columns (representing the different IR techniques). The scores in each row are ranked from 1 to k, and the Friedman test determines the probability that the rank totals of the different columns differ significantly from what could be expected, if the data had occurred entirely by chance. (Siegel and Castellan, 1988, p. 175-176)

The batch mode module of the SPSS statistical software for Windows was used to conduct all tests. It uses the following formula to calculate the test statistic, which allows for ties in the data (i.e., that two IR techniques perform equally for a given topic)[61]:

$$F_r = \frac{\left(12 \Big/ Nk(k+1)\sum_{l=1}^{k} C_l^2 - 3N(k+1)\right)}{1 - \sum T \Big/ Nk\left(k^2 - 1\right)} \qquad (9)$$

where

$$\sum T = \sum_{i=1}^{N}\sum_{l=1}^{k}\left(t^3 - t\right)$$

$N$ = the number of topics,
$k$ = the number of IR techniques under comparison,
$C$ = sum of ranks in the $l^{th}$ column (i.e., the sum of ranks for the $l^{th}$ IR technique),
$t$ = the number of sets of tied ranks for each topic $i$.
(Siegel and Castellan, 1988, p. 176-180; [SPPS Inc.], [2002])

The null hypothesis $H_0$ is that the data resulting from the $k$ IR techniques all come from a populations with the same median, i.e., that they do not differ in IR performance. If the Friedman test allows rejection of the null hypothesis the tests indicate that *at least one* of the techniques differs from *at least one* of the other techniques, but not which one is different, or how many of the techniques are different from each other. Two methods, $R_u$ and $R_v$, are known to differ if the following inequality is satisfied:

[61] A more powerful version of the Friedman test is given in Hull (Hull, 1993). This is unfortunately not offered by SPSS.

$$\left|R_u - R_v\right| \geq z_{\alpha/k(k-1)}\sqrt{\frac{Nk(k+1)}{6}} \tag{10}$$

where $R_u$ and $R_v$ are the rank sums of the $u^{th}$ and $v^{th}$ IR techniques,
$N$ and $k$ as given above, and
$z_{\alpha/k(k-1)}$ gives the critical value of $z$ above which lies the $\alpha/k(k-1)$ percent of the normal distribution (Siegel and Castellan, 1988, p. 180-181).

All significance tests in the dissertation was conducted at $\alpha = 0.05$ (also referred to as the p-value).

## 6.6 Summary statements

The chapter has described the test data and methods use in the main experiment as a background for the results in Chapter 7, and the discussion in Chapter 8. Accounts have been given of the INEX test collection, the IR systems and test databases used including the citation index created for the experiment. In addition each of the test runs have been described, as have the performance measures and statistical tests used.

# 7 Results

In this chapter the data collected from the main experiment are analysed and the results described. The main discussion of the results is placed in Chapter 7 because there seems to be interactions between several of the research questions.

First, a brief account is given of the runs submitted to INEX 2002 and the experiences gained from carrying out these runs. This is followed by the results from the main experiments: Section 7.2.1 contains descriptive statistics on different aspects of the best match boomerang effect, Section 7.2.2 presents the overall results tables and describes the general trends, and Section 7.2.3 deals with the research questions one by one including the results of the statistical tests. The chapter ends with summary statements in Section 7.3.

All results presented in this chapter have been calculated using the relevance assessment package version 1.8, and `inex_eval` version 0.007 modified to include the str_whole and gen_whole quantification functions. All P-R and (nD)CG curves are reproduced in larger scale in Appendix 8 and 9 to facilitate more detailed interpretations.

## 7.1 The official INEX 2002 runs

A total of 49 CO runs were submitted by 24 participants. Overall, the result of INEX 2002 displayed rather low levels of average precision (AvgP) for the CO topics: the highest gen_inex value was 0.0705, and the highest str_inex value was 0.0883 for all 49 runs (Gövert and Kazai, 2003)[62].

Table 7.1 presents an overview of the results of the three preliminary runs submitted to INEX 2002 as part of the TAPIR project. The official results of the runs are reproduced in entirety in Appendix 7. The best match boomerang effect run submitted to INEX

[62] The CAS topics displayed significantly higher values: the highest str_inex value was 0.3438, and the highest gen_inex value was 0.2752 (Gövert and Kazai, 2003).

2002 used the extended citation indexes both for the extraction of citations for Step 2 and as the index against which the weighted citation queries were matched in Step 3. All citations from Step 2 were used in the forward chaining, that is, no CCV_step2 threshold was invoked. The boomerang and the polyrepresentation baseline were run with a fixed DCV_step1 of 500 documents. However, the greatest difference to the main experiment presented below is that the Description element of the INEX CO topics were used as part of the queries in addition to the Title and Keywords elements. Only the two latter were used in the main experiment.

**Table 7.1. AvgP scores for the threes official runs submitted to INEX2002. Values in bold face denote the best run in each column.**

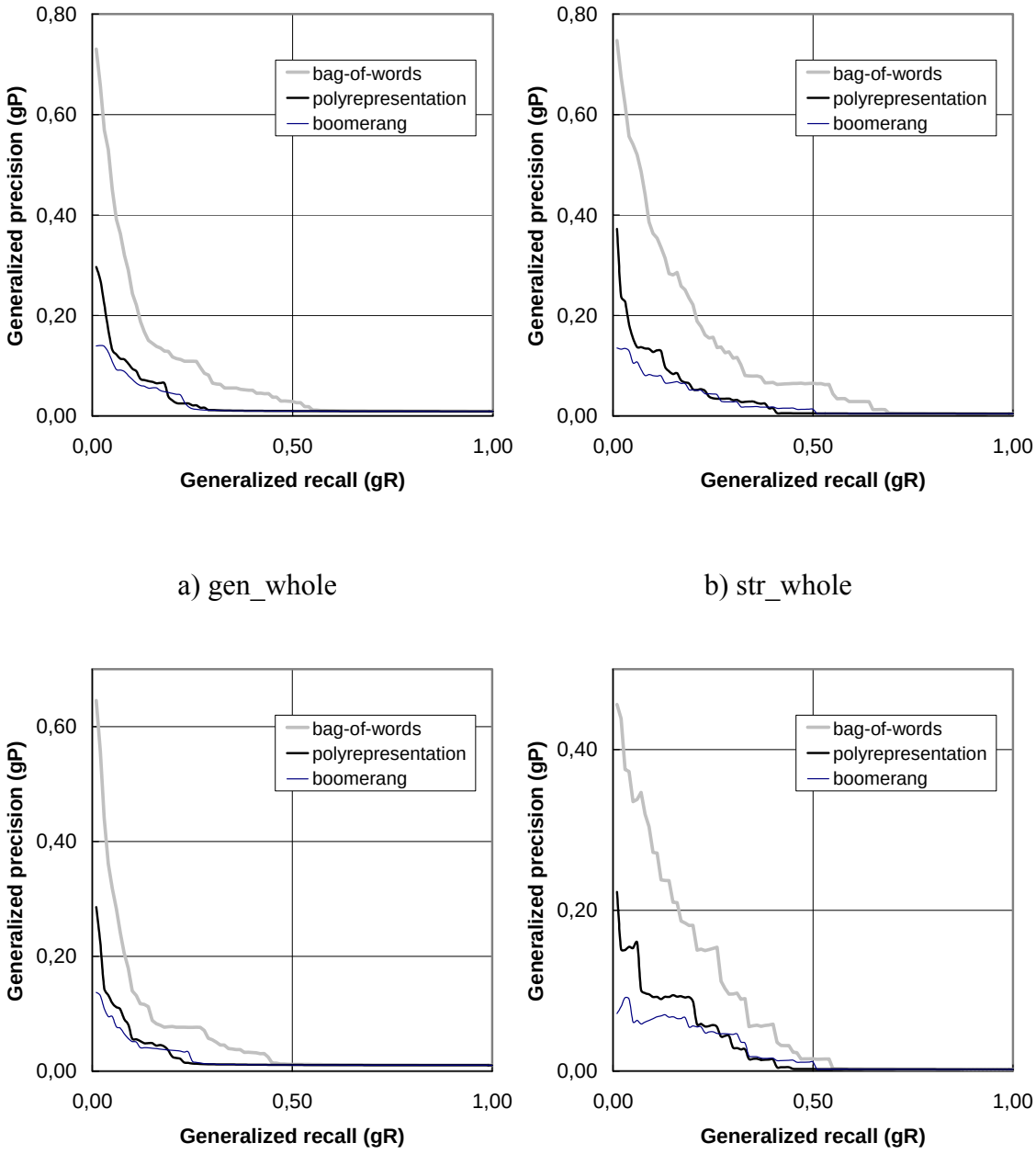
| INEX 2002 run name | gen_whole | str_whole | gen_inex | str_inex |
|---|---|---|---|---|
| boomerang | 0.0250 | 0.0270 | 0.0227 | 0.0231 |
| polyrepresentation | 0.0316 | 0.0365 | 0.0271 | 0.0313 |
| bag-of-words | **0.0853** | **0.1188** | **0.0618** | **0.0809** |
| Statistical significance (p = 0.05): | bag-of-words > polyrepresentation, boomerang | | | |

It can be observed that the bag-of-words baseline displays a significantly higher level of AvgP over all four quantification functions compared to the polyrepresentation baseline and the boomerang run. The bag-of-words baseline also displayed high performance in absolute terms in comparison to the other official submissions: it ranked 3rd (gen_inex) and 2nd (str_inex) in the INEX 2002 top 10 (Gövert and Kazai, 2003, p. 15). The boomerang run was at rank 37 and 33, and the polyrepresentation baseline at rank 34 and 32 for the gen_inex and str_inex functions respectively in the 49 submissions.

From Appendix 7 it can be seen that the bag-of-words baseline is very close to having the best performance along the top of all the P-R curves in INEX 2002, and that it does better than the median performance in almost every topic. In contrast, both the boomerang run and the polyrepresentation baseline are not near the top of the P-R curves at any point. In particular, the boomerang effect has problems with achieving high precision even at low recall levels. In Figure 7.1 all three runs are displayed within the same P-R curve for the gen_whole and str_whole functions (curves a and b), and for INEX's standard quantification functions (curves c and d). At all levels of recall the bag-of-words baseline lies above the other two. The polyrepresentation baseline displays higher AvgP than the boomerang run at low recall levels, but the boomerang

run is *on par* or displays slightly higher AvgP at different places in the 0.2 - 0.5 recall interval depending on the quantification function. There is, however, no overall statistical difference between the two runs. Although the same tendencies are apparent in all four curves the differences are more obvious with the strict functions.

It is not possible to deduce *why* the bag-of-words baseline performed so much better than the other two runs in INEX 2002 solely by looking at the results. Instead the rather discouraging results in INEX2002 with the best match boomerang effect and the polyrepresentation baseline made us attempt to identify factors that affect the performance of these two IR techniques as described in Chapter 6. The DCV_step1 and CCV_step2 variables, as well as the flat and expanded citation indexes, were implemented in the main experiment in order to gain further knowledge of what affects the performance of the best match boomerang effect and the polyrepresentation baseline.




a) gen_whole

b) str_whole

c) gen_inex

d) str_inex

**Figure 7.1.a-d. P-R curves of the three official runs submitted to INEX 2002 for the four quantification functions. Note that the y-axis is fitted to each curve.**

## 7.2 Main experiment runs

### *7.2.1 Characteristics of the best match boomerang effect*

As in the pre-experiment it is important to know whether the overall framework supports the conducted runs. For instance, if only very few documents can be identified in Step 1, neither the boomerang effect nor the polyrepresentation baseline have much chance of performing well. This section reports some of the statistics that were collected during the execution of the runs.

After the parsing of the topics in InQuery there were 9.0 unique query keys per topic on average in Step 1, and 12.5 in total including those with a frequency of more than 1. These figures are for the query keys after stemming and removal of stop words by InQuery as described in Section 6.3.1 above. Because the same queries were used in all three run types the figures cover all runs in the main experiment. Table 7.2 shows an overview of the number of documents retrieved as result of the queries in Step 1 of the best match boomerang effect when the DCV_step1 threshold not is invoked (These were also the runs used as sources for the polyrepresentation baseline). The average number of retrieved documents per representation is quite close to what could be expected when it is compared to the number of unique index keys per representation reported in Table 6.3: the same four representations are in the top (CTI, INT, FGC, ABS) and bottom (TBC, AKW, ATL, DE) in both tables. This is quite natural as a query will have a greater chance of matching more documents in representations which have a large number of unique index keys. This leaves the question of whether enough documents were retrieved in all representations to support the extraction of references in the best match boomerang effect. The largest value of DCV_step1 that was tested was 512. It can be seen that only the table captions representation retrieved less that 512 documents per topic on average. However, it is obvious from the standard deviation, the minimum and the maximum values in Table 7.2 that the averages conceal large variation in individual queries. All representations have relatively large standard deviations from the average, and very few documents were retrieved in some topics. Table 7.3 shows the number of topics for each representation that retrieved *fewer* documents than a given DCV_step1 threshold. The table captions and author keywords account for more than half of these at DCV_step1 = 512 (11+16/51). Except for these two representations, the number of topics with fewer documents retrieved does not seem

to be detrimental for the functioning of the best match boomerang effect. Indeed, the variation across representations is an inescapable consequence of the uncertainties and inconsistencies inherent in IR as predicted by the theory of polyrepresentation, but it needs to be kept in mind when interpreting the results of the main experiment. Note that Table 7.2 shows the number of documents retrieved in InQuery, not if any of them are relevant to the topics. Since no structure is imposed on the queries sent to InQuery, a document will be retrieved in Step 1 if it is indexed with just one of the query keys. Performance results for the ten runs in the table are presented in Section 7.2.3.2 below.

**Table 7.2. Number of documents identified in Step 1 in the best match boomerang effect when the DCV_step1 threshold is not invoked. The average number of documents retrieved over 24 topics is shown as well as the minimum, the maximum, and the standard deviation.**

| Representation | Average | St.dev | Min. | Max. |
|---|---|---|---|---|
| Article title (ATL) | 825 | 468 | 38 | 1,716 |
| Abstract (ABS) | 2,195 | 1,179 | 103 | 4,398 |
| Author keywords (AKW) | 783 | 988 | 3 | 3,759 |
| Figure captions (FGC) | 2,244 | 1,108 | 94 | 3,913 |
| Table captions (TBC) | 398 | 287 | 13 | 956 |
| Introductions (INT) | 2,326 | 1,017 | 225 | 3,651 |
| Conclusions (CON) | 1,217 | 673 | 74 | 2,402 |
| Cited titles (CTI) | 3,095 | 1,553 | 42 | 5,338 |
| Descriptors (DE) | 1,077 | 650 | 1 | 2,451 |
| Identifiers (ID) | 1,699 | 874 | 31 | 3,108 |

**Table 7.3. Number of topics out of 24 in each representation that retrieved fewer documents in Step 1 of the best match boomerang effect than the DCV_step1 threshold.**

| | DCV_step1 | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Representation | 2 | 4 | 8 | 16 | 32 | 64 | 128 | 256 | 512 |
| Article title (ATL) | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 3 | 7 |
| Abstract (ABS) | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 2 |
| Author keywords (AKW) | 0 | 1 | 1 | 1 | 2 | 2 | 4 | 8 | 11 |
| Figure captions (FGC) | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 2 |
| Table captions (TBC) | 0 | 0 | 0 | 1 | 1 | 3 | 5 | 10 | 16 |
| Introductions (INT) | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 2 |
| Conclusions (CON) | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 2 | 4 |
| Cited titles (CTI) | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 1 |
| Descriptors (DE) | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 2 | 4 |
| Identifiers (ID) | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 1 | 2 |
| Sum | 1 | 2 | 2 | 3 | 5 | 9 | 17 | 30 | 51 |

Table 7.4 shows the average number of seed documents used per topic in Step 3 in the best match boomerang effect distributed on DCV_step1 values, and over CCV_step2 values. The Max values denote the number of citations if the CCV_step2 threshold had not been invoked, and the Low, Medium and High thresholds are implemented as percentages of this, as described in Section 6.3.2. It can be seen that the average number of seed documents per topic rises in a steady exponential function over DCV_step1 values[63], which is not entirely surprising considering that the DCV_step1 values are chosen from an exponential function. In light of the irregularities discussed above it is, however, interesting that the data fit so well on an exponential curve. It may also be observed that a fairly large amount of seed documents are available for the forward chaining. For instance, even at DCV_step1 = 2 there are no less than 77 seed documents available per topic on average if no CCV_step2 threshold were to be invoked. Note that there is no difference in the number of citations between the flat and the expanded citation index because the same number of unique citations is extracted: Only the *frequencies* that determine the weights are different. Thus, although the number of

[63] In fact, an exponential trend line can be fitted to the data with $R^2 = 0.999$.

citations will remain the same, the actual citations that are selected and their weights may be different with the flat or the expanded citation index.

**Table 7.4. Average number of seed documents per topic in Step 3 in the best match boomerang effect distributed on DCV_step1 values, and over CCV_step2 values. The Max values denote the number of citations if the CCV_step2 threshold had not been invoked.**

| | DCV_step1 | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| CCV_step2 | 2 | 4 | 8 | 16 | 32 | 64 | 128 | 256 | 512 |
| Low (25 %) | 19 | 32 | 60 | 118 | 232 | 451 | 907 | 1,822 | 3,495 |
| Medium (50 %) | 39 | 65 | 120 | 236 | 463 | 902 | 1,814 | 3,645 | 6,990 |
| High (75 %) | 58 | 97 | 180 | 353 | 695 | 1,353 | 2,721 | 5,467 | 10,486 |
| Max (100 %) | 77 | 129 | 240 | 471 | 927 | 1,804 | 3,628 | 7,289 | 13,981 |

Table 7.5 shows the average number of documents retrieved by the polyrepresentation baseline and in Step 3 of the best match boomerang effect. The data is distributed on DCV_step1 values for both, and over CCV_step2 values, and on whether a flat or extended citation index was used for the extraction of citations for Step 2 for the best match boomerang effect. The Max values denote the number of documents retrieved if the CCV_step2 threshold had not been invoked. Note that no more than 100 documents were retrieved even when possible, because the DCV_step3 threshold was set to a maximum of 100 documents in all runs as discussed in Section 6.3 above. It can be observed that both the polyrepresentation baseline and the best match boomerang effect retrieved at least 100 documents on average per topic at high values of DCV_step1. As expected the polyrepresentation baseline retrieved less than 100 documents on average for low values of DCV_step1 since there were not enough documents in each of the ten representations to add up to 100. The maximum number of documents that in theory can be retrieved by the polyrepresentation baseline at a given value of DCV_step1 can be calculated by multiplying the value of DCV_step1 with the number of representations. In theory, 100 documents cannot be retrieved for values of DCV_step1 of less than 10 (10 documents × 10 representations = 100). Because fewer documents than DCV_step1 are retrieved for some topics in several representations (See Table 7.2 above), and because of *overlaps* among the documents retrieved for a topic in different representations, the number of documents retrieved for the polyrepresentation baseline is significantly lower than the number of documents that could have been retrieved in theory. It is interesting to note that the number of documents retrieved by the polyrepresentation baseline remains fairly constant at the double of the DCV_step1

value until the threshold of 100 documents is met, which happens when DCV_step1 is between 32 and 64. This indicates that the proportion of documents in overlaps is roughly the same on average in this interval. The exponential increase is also found in the number of documents retrieved by the polyrepresentation baseline ($R^2$ = 0.999). Again it is interesting that an exponential function can be fitted so well to the data considering the irregularities discussed above. Although it could to a certain extent be expected, it is not known exactly what causes the data to fit so well instead of deviating from an exponential function.

**Table 7.5. The average number of documents retrieved per topic by the polyrepresentation baseline and in Step 3 of the best match boomerang effect. The data is distributed on DCV_step1 values, over CCV_step2 values, and on whether a flat (f) or extended (x) citation index was used for the extraction of citations for Step 2. The Max values denote the number of documents retrieved on average if the CCV_step2 threshold had not been invoked.**

| | | DCV_step1 | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| CCV_step2 | CI | 2 | 4 | 8 | 16 | 32 | 64 | 128 | 256 | 512 |
| Low | f | 35 | 53 | 67 | 88 | 98 | 100 | 100 | 100 | 100 |
| Medium | f | 53 | 67 | 83 | 99 | 100 | 100 | 100 | 100 | 100 |
| High | f | 58 | 75 | 92 | 100 | 100 | 100 | 100 | 100 | 100 |
| Max | f | 65 | 79 | 96 | 100 | 100 | 100 | 100 | 100 | 100 |
| Low | x | 32 | 49 | 72 | 94 | 100 | 100 | 100 | 100 | 100 |
| Medium | x | 50 | 66 | 88 | 99 | 100 | 100 | 100 | 100 | 100 |
| High | x | 60 | 75 | 92 | 100 | 100 | 100 | 100 | 100 | 100 |
| Max | x | 65 | 79 | 96 | 100 | 100 | 100 | 100 | 100 | 100 |
| Polyrepresentation | - | 4 | 8 | 17 | 34 | 71 | 100 | 100 | 100 | 100 |

In contrast, the best match boomerang effect could in theory retrieve at least 100 documents for every topic at all levels of DCV_step1. This could happen if documents retrieved from different representations contained the same reference, and if this reference was among the most cited in the corpus. In practice, fewer than 100 documents were retrieved by the best match boomerang effect at low values of DCV_step1 in the main experiment. Without the CCV_step2 threshold (the Max values in Table 7.5) 100 documents were retrieved on average at DCV_step1 = 16. The use of the CCV_step2 threshold to reduce the number of citations in the weighted sum query also reduced the number of documents retrieved in Step 3. However, the number of

documents retrieved was not reduced proportionally to the reduction in the number of citations. This indicates that the best match boomerang effect was successful in identifying seed documents in the top of the weighted query that were cited in several documents. Note that there were small differences in the number of retrieved documents when the CCV_step2 threshold was used, depending on whether a flat or an expanded citation index was used for the extraction of citations to Step 2. This can be explained by the fact that different citations were selected as seed documents for the weighted sum queries because of the use of either the flat or the expanded citation index. Performance differences between the use of two types of citation indexes are reported in Section 7.2.3.6 and Section 7.2.3.7 below.

### *7.2.2 Overall trends in the precision of the runs*

An overview of the results of the main experiment is presented in the form of tables that report the AvgP of the runs carried out across the tested variables. Table 7.6 to Table 7.9 below each presents the AvgP of 108 best match boomerang effect runs, nine polyrepresentation baseline runs and one bag-of-words baseline run[64] for each of the four quantification functions: gen_whole, str_whole, gen_inex and str_inex. In these four tables values in **bold** denote the best run in each row (where each row covers the whole range of DCV_step1 values for a particular combination of CCV_step1 and the flat or extended citation indexes). For the best match boomerang effect runs shaded values denote the best combination of flat and expanded citation indexes for a given value of DCV_step1 and CCV_step2. The boxed run points to the best match boomerang effect run with the best AvgP value in the table. Note that the absolute values cannot be compared across the tables since the quantification functions are different. The AvgP of the runs carried out with each of the ten functional and cognitive representations is presented separately in Table 7.10 for all four quantification functions. The data behind these five tables is used for the statistical testing of the research questions below. First, a number of general trends of the absolute AvgP values in the tables are described:

---

[64] Although no DCV_step1 threshold is associated with the bag-of-words baseline it was chosen to present the results of it in the same table as the best match boomerang effect runs and the polyrepresentation baseline, as it belongs together with these. It has been placed between DCV_step1 64 and 128 to indicate that the top 100 documents were evaluated for each run.

- Compared with the performance of the official runs submitted to INEX 2002 (Table 7.1) the performance of the best match boomerang effect and the polyrepresentation baseline improved substantially when the best possible combinations of variables in the main experiment were taken into account. For instance, the best match boomerang effect improved from 0.0270 to 0.0718 AvgP, and the polyrepresentation baseline from 0.0365 to 0.0799 for the str_whole quantification function (Table 7.7). However, the bag-of-words baseline still displayed the highest AvgP of all runs regardless of quantification function[65].
- The best performing polyrepresentation baseline run was always at DCV_step1 = 64 for all quantification functions. As discussed above, although a full set of 100 documents might, in theory, be retrieved at lower values of DCV_step1, in practice this did not happen until DCV_step1 = 64 on average (See Table 7.5 above). Maximum AvgP of the polyrepresentation baseline is hence not to be expected at values lower than 64. For the DCV_step1 values greater than 64 that were tested in the main experiment, the AvgP drops gradually for the polyrepresentation baseline across all quantification functions.
- For any specific combination of CCV_step2 and the citation indexes (i.e., a row in the tables) the best match boomerang effect achieves the maximum AvgP scores at relatively low values of DCV_step1 and within certain ranges of it: for the two generalized quantification functions all the runs with highest AvgP occur at DCV_step1 = 16 or 32, and for the str_whole function at DCV_step1 = 8 or 16. Measured with the str_inex function the maximum AvgP is reached at DCV_step1 = 8 except for two rows (which reach maximum at DCV_step1 = 2 and 4).
- When considering the best performing runs in each row as measured by the gen_whole, str_whole and partly the gen_inex quantification functions, a pattern emerges: Runs where citations have been extracted for Step 2 from the expanded citation index (the xf and xx runs in the tables) most often achieve the best AvgP

[65] The values for the bag-of-words baseline differ slightly from the ones presented in section 7.1 and Appendix 7. This is due to the correction of a number of erroneous document IDs that were corrected after the submission to INEX 2002.

score at a lower value of DCV_step1, compared to when citations were extracted from a flat citation index (the ff and fx runs in the tables). For the str_inex quantification function there were no differences between the two types of citation indexes except for two runs, which achieved the best results at a lower DCV_step1 with either the ff or fx run.

- There is a general tendency for the absolute AvgP scores to increase over the three values of CCV_step1. For the gen_whole, str_whole, and gen_inex quantification functions this occurred in 89 % of the runs from Low to Medium and Low to High, and in 67-69 % from Medium to High. For the str_inex quantification function the corresponding values were lower: 69 % from Low to Medium, 78 % from Low to high, and 64 % from Medium to High. For the gen_whole, str_whole and gen_inex functions the main part of the AvgP that showed the opposite behaviour (i.e., decreased) was found at values of DCV_step1 from 128 and upwards. The cases with a decrease was more spread out when measured with the str_inex function – beginning already at DCV_step1 = 8.
- Of the four variables tested for the best match boomerang effect (i.e., DCV_step1, CCV_step2, and the use of flat or expanded citation indexes for Step 2 and Step 3), the DCV_step1 variable appears to be the determining variable. The variation over this variable is much greater than over the other three in absolute AvgP values. The general tendency across the range of DCV_step1 values is that the AvgP scores first rise from the level achieved at low values of DCV_step1 to their maximum, and then drop again to reach the lowest AvgP scores at high values of DCV_step1. This pattern can be found in almost all combinations of CCV_step2 and the citation indexes with all four quantification functions.
- As an individual representation the cited titles (CTI, see Table 7.10) do remarkably well compared to the results achieved by Salton and Zhang (1986). For the gen_whole and str_whole quantification functions the cited titles obtain greater AvgP scores than all best match boomerang effect and polyrepresentation baseline runs (See Table 7.6 and Table 7.7). When measured with the str_whole quantification function the cited titles perform better than the best match boomerang effect, but not better than the polyrepresentation baseline. For the str_inex quantification function, however, the polyrepresentation

baseline and several of the best match boomerang effect runs achieve better AvgP scores than the cited titles.

- Apart from the cited titles, the abstracts, identifiers, and article titles achieved the highest AvgP scores of the individual representations in Table 7.10. The table captions have the lowest score in all quantification functions. This is followed by the conclusions, figure captions and author keywords which are also among the lowest performing in the gen_whole, str_whole, and gen_inex quantification functions. Apart from the table captions, low AvgP scores are obtained by the conclusions and the descriptors in the str_inex quantification function.
- Overall the gen_whole, str_whole, and gen_inex quantification functions tend to yield similar results when the absolute AvgP scores are ignored and only the ranking order of the runs are considered. The str_inex quantification function often deviates from the other three.

**Table 7.6. Average precision values (quantification function: gen_whole) for the two baseline runs and the best match boomerang effect runs. Values are given for three variables for the boomerang effect: DCV_step1 (2-512 documents), CCV_step2 (Low, Medium, and High thresholds), and flat or expanded Citation Indexes in either Step 2 and Step 3 (ff, fx, xf, or xx). For the bag-of-words baseline values are given for DCV_step1. Values in bold face denote the best run in each row, shaded values denote the best run for a given combination of DCV_step1 and CCV_step2, and the boxed run is the best of the best match boomerang effect runs.**

| | | DCV_step1 | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| CCV | CI | 2 | 4 | 8 | 16 | 32 | 64 | 128 | 256 | 512 |
| L | ff | 0.0380 | 0.0404 | 0.0452 | 0.0483 | **0.0503** | 0.0486 | 0.0439 | 0.0366 | 0.0308 |
| L | fx | 0.0381 | 0.0405 | 0.0450 | 0.0482 | **0.0487** | 0.0476 | 0.0421 | 0.0350 | 0.0299 |
| L | xf | 0.0350 | 0.0402 | 0.0457 | **0.0468** | 0.0458 | 0.0424 | 0.0362 | 0.0314 | 0.0265 |
| L | xx | 0.0352 | 0.0399 | 0.0445 | **0.0466** | 0.0454 | 0.0399 | 0.0348 | 0.0301 | 0.0250 |
| M | ff | 0.0403 | 0.0445 | 0.0488 | 0.0523 | **0.0532** | 0.0505 | 0.0449 | 0.0360 | 0.0306 |
| M | fx | 0.0400 | 0.0443 | 0.0487 | **0.0516** | 0.0514 | 0.0491 | 0.0432 | 0.0346 | 0.0297 |
| M | xf | 0.0394 | 0.0432 | 0.0485 | **0.0491** | 0.0485 | 0.0446 | 0.0370 | 0.0320 | 0.0270 |
| M | xx | 0.0391 | 0.0432 | 0.0470 | **0.0487** | 0.0480 | 0.0413 | 0.0355 | 0.0303 | 0.0254 |
| H | ff | 0.0415 | 0.0465 | 0.0515 | 0.0542 | **0.0545** | 0.0507 | 0.0445 | 0.0359 | 0.0304 |
| H | fx | 0.0416 | 0.0459 | 0.0509 | **0.0531** | 0.0523 | 0.0491 | 0.0429 | 0.0345 | 0.0294 |
| H | xf | 0.0401 | 0.0460 | 0.0494 | **0.0519** | 0.0496 | 0.0449 | 0.0372 | 0.0318 | 0.0269 |
| H | xx | 0.0401 | 0.0455 | 0.0486 | **0.0508** | 0.0490 | 0.0416 | 0.0354 | 0.0302 | 0.0253 |
| Polyrep. | | 0.0193 | 0.0266 | 0.0352 | 0.0466 | 0.0541 | **0.0554** | 0.0469 | 0.0441 | 0.0436 |
| Bag-of-words | | | | | | | **0.0837** | | | |

**Table 7.7. Average precision values (quantification function: str_whole) for the two baseline runs and the best match boomerang effect runs. Values are given for three variables for the boomerang effect: DCV_step1 (2-512 documents), CCV_step2 (Low, Medium, and High thresholds), and flat or expanded Citation Indexes in either Step 2 and Step 3 (ff, fx, xf, or xx). For the bag-of-words baseline values are given for DCV_step1. Values in bold face denote the best run in each row, shaded values denote the best run for a given combination of DCV_step1 and CCV_step2, and the boxed run is the best of the best match boomerang effect runs.**

| | | DCV_step1 | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| CCV | CI | 2 | 4 | 8 | 16 | 32 | 64 | 128 | 256 | 512 |
| L | ff | 0.0519 | 0.0536 | 0.0632 | **0.0653** | 0.0627 | 0.0585 | 0.0502 | 0.0390 | 0.0326 |
| L | fx | 0.0525 | 0.0546 | 0.0630 | **0.0654** | 0.0609 | 0.0573 | 0.0481 | 0.0375 | 0.0322 |
| L | xf | 0.0510 | 0.0576 | **0.0674** | 0.0634 | 0.0567 | 0.0515 | 0.0409 | 0.0341 | 0.0293 |
| L | xx | 0.0519 | 0.0580 | **0.0660** | 0.0648 | 0.0595 | 0.0488 | 0.0405 | 0.0340 | 0.0284 |
| M | ff | 0.0559 | 0.0603 | 0.0671 | **0.0693** | 0.0667 | 0.0610 | 0.0518 | 0.0381 | 0.0323 |
| M | fx | 0.0546 | 0.0617 | 0.0666 | **0.0685** | 0.0639 | 0.0592 | 0.0494 | 0.0372 | 0.0321 |
| M | xf | 0.0580 | 0.0622 | **0.0695** | 0.0667 | 0.0622 | 0.0543 | 0.0418 | 0.0347 | 0.0300 |
| M | xx | 0.0588 | 0.0622 | **0.0682** | 0.0671 | 0.0622 | 0.0500 | 0.0415 | 0.0342 | 0.0286 |
| H | ff | 0.0595 | 0.0644 | 0.0697 | **0.0718** | 0.0688 | 0.0610 | 0.0510 | 0.0382 | 0.0322 |
| H | fx | 0.0602 | 0.0645 | 0.0689 | **0.0703** | 0.0657 | 0.0588 | 0.0488 | 0.0372 | 0.0316 |
| H | xf | 0.0593 | 0.0659 | **0.0715** | 0.0695 | 0.0638 | 0.0546 | 0.0420 | 0.0346 | 0.0299 |
| H | xx | 0.0600 | 0.0653 | **0.0708** | 0.0692 | 0.0636 | 0.0502 | 0.0413 | 0.0340 | 0.0285 |
| Polyrep. | | 0.0327 | 0.0439 | 0.0569 | 0.0708 | 0.0783 | **0.0799** | 0.0710 | 0.0674 | 0.0663 |
| Bag-of-words | | | | | | | **0.1193** | | | |

**Table 7.8. Average precision values (quantification function: gen_inex ) for the two baseline runs and the best match boomerang effect runs. Values are given for three variables for the boomerang effect: DCV_step1 (2-512 documents), CCV_step2 (Low, Medium, and High thresholds), and flat or expanded Citation Indexes in either Step 2 and Step 3 (ff, fx, xf, or xx). For the bag-of-words baseline values are given for DCV_step1. Values in bold face denote the best run in each row, shaded values denote the best run for a given combination of DCV_step1 and CCV_step2, and the boxed run is the best of the best match boomerang effect runs.**

| | | DCV_step1 | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| CCV | CI | 2 | 4 | 8 | 16 | 32 | 64 | 128 | 256 | 512 |
| L | ff | 0.0301 | 0.0317 | 0.0354 | 0.0378 | **0.0393** | 0.0382 | 0.0354 | 0.0313 | 0.0270 |
| L | fx | 0.0303 | 0.0322 | 0.0352 | 0.0374 | **0.0382** | 0.0373 | 0.0343 | 0.0300 | 0.0263 |
| L | xf | 0.0283 | 0.0320 | 0.0358 | **0.0365** | 0.0361 | 0.0343 | 0.0304 | 0.0273 | 0.0237 |
| L | xx | 0.0286 | 0.0321 | 0.0354 | **0.0361** | 0.0359 | 0.0328 | 0.0300 | 0.0263 | 0.0227 |
| M | ff | 0.0319 | 0.0347 | 0.0380 | 0.0402 | **0.0410** | 0.0393 | 0.0362 | 0.0309 | 0.0269 |
| M | fx | 0.0319 | 0.0350 | 0.0379 | 0.0396 | **0.0400** | 0.0385 | 0.0348 | 0.0299 | 0.0261 |
| M | xf | 0.0315 | 0.0341 | 0.0378 | 0.0379 | **0.0383** | 0.0358 | 0.0311 | 0.0277 | 0.0241 |
| M | xx | 0.0314 | 0.0346 | 0.0367 | 0.0379 | **0.0381** | 0.0340 | 0.0305 | 0.0265 | 0.0231 |
| H | ff | 0.0330 | 0.0361 | 0.0395 | 0.0415 | **0.0422** | 0.0395 | 0.0359 | 0.0308 | 0.0269 |
| H | fx | 0.0332 | 0.0358 | 0.0389 | **0.0408** | 0.0407 | 0.0384 | 0.0347 | 0.0298 | 0.0259 |
| H | xf | 0.0322 | 0.0356 | 0.0383 | **0.0401** | 0.0388 | 0.0361 | 0.0312 | 0.0276 | 0.0241 |
| H | xx | 0.0322 | 0.0356 | 0.0379 | **0.0392** | 0.0386 | 0.0342 | 0.0304 | 0.0265 | 0.0230 |
| Polyrep. | | 0.0161 | 0.0202 | 0.0257 | 0.0349 | 0.0409 | **0.0419** | 0.0355 | 0.0335 | 0.0332 |
| Bag-of-words | | | | | | | **0.0606** | | | |

**Table 7.9. Average precision values (quantification function: str_inex) for the two baseline runs and the best match boomerang effect runs. Values are given for three variables for the boomerang effect: DCV_step1 (2-512 documents), CCV_step2 (Low, Medium, and High thresholds), and flat or expanded Citation Indexes in either Step 2 and Step 3 (ff, fx, xf, or xx). For the bag-of-words baseline values are given for DCV_step1. Values in bold face denote the best run in each row, shaded values denote the best run for a given combination of DCV_step1 and CCV_step2, and the boxed run is the best of the best match boomerang effect runs.**

| | | DCV_step1 | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| CCV | CI | 2 | 4 | 8 | 16 | 32 | 64 | 128 | 256 | 512 |
| L | ff | 0.0486 | 0.0483 | **0.0561** | 0.0557 | 0.0509 | 0.0409 | 0.0352 | 0.0307 | 0.0257 |
| L | fx | 0.0492 | 0.0493 | **0.0556** | 0.0550 | 0.0476 | 0.0398 | 0.0341 | 0.0292 | 0,0248 |
| L | xf | 0.0516 | 0.0555 | **0.0657** | 0.0544 | 0.0493 | 0.0427 | 0.0358 | 0.0294 | 0.0239 |
| L | xx | 0.0533 | 0.0579 | **0.0636** | 0.0568 | 0.0543 | 0.0419 | 0.0358 | 0.0295 | 0.0236 |
| M | ff | 0.0550 | 0.0564 | **0.0606** | 0.0549 | 0.0501 | 0.0422 | 0.0363 | 0.0306 | 0.0254 |
| M | fx | 0.0526 | 0.0587 | **0.0604** | 0.0539 | 0.0466 | 0.0406 | 0.0351 | 0.0294 | 0.0247 |
| M | xf | 0.0599 | 0.0589 | **0.0655** | 0.0568 | 0.0534 | 0.0429 | 0.0354 | 0.0297 | 0.0247 |
| M | xx | 0.0613 | 0.0599 | **0.0641** | 0.0579 | 0.0544 | 0.0410 | 0.0360 | 0.0294 | 0.0240 |
| H | ff | 0.0599 | **0.0608** | 0.0601 | 0.0566 | 0.0516 | 0.0423 | 0.0362 | 0.0309 | 0.0253 |
| H | fx | **0.0607** | 0.0607 | 0.0596 | 0.0549 | 0.0474 | 0.0403 | 0.0353 | 0.0295 | 0.0242 |
| H | xf | 0.0606 | 0.0625 | **0.0675** | 0.0570 | 0.0540 | 0.0428 | 0.0355 | 0.0296 | 0.0246 |
| H | xx | 0.0626 | 0.0618 | **0.0652** | 0.0583 | 0.0549 | 0.0408 | 0.0358 | 0.0293 | 0.0239 |
| Polyrep. | | 0.0432 | 0.0474 | 0.0567 | 0.0621 | 0.0675 | **0.0678** | 0.0624 | 0.0616 | 0.0616 |
| Bag-of-words | | | | | | | **0.0830** | | | |

**Table 7.10. Average precision values for natural language queries run against the individual representations. Values are given for each of the four quantification functions. Values in bold face denote the best run in the column.**

| Representation | gen_whole | str_whole | gen_inex | str_inex |
|---|---|---|---|---|
| Article title (ATL) | 0.0422 | 0.0586 | 0.0344 | 0.0516 |
| Abstract (ABS) | 0.0424 | 0.0645 | 0.0324 | **0.0604** |
| Author keywords (AKW) | 0.0269 | 0.0303 | 0.0240 | 0.0366 |
| Figure captions (FGC) | 0.0261 | 0.0403 | 0.0212 | 0.0415 |
| Table captions (TBC) | 0.0123 | 0.0103 | 0.0118 | 0.0074 |
| Introductions (INT) | 0.0336 | 0.0429 | 0.0275 | 0.0439 |
| Conclusions (CON) | 0.0250 | 0.0289 | 0.0219 | 0.0264 |
| Cited titles (CTI) | **0.0605** | **0.0762** | **0.0466** | 0.0553 |
| Descriptors (DE) | 0.0391 | 0.0434 | 0.0311 | 0.0278 |
| Identifiers (ID) | 0.0485 | 0.0629 | 0.0382 | 0.0503 |

### *7.2.3 Research questions*

#### 7.2.3.1 Research question 1: Does the best match boomerang effect achieve a similar level of performance compared to what is obtainable with a polyrepresentation baseline and a best match baseline?

The question investigates whether the best match boomerang effect can achieve a similar level of performance as that of the bag-of-words baseline and the polyrepresentation baseline. In order to maintain comparability across variables, a specific combination of CCV_step2 and citations indexes was chosen for each run to represent the best match boomerang effect: The combination of expanded citation indexes and a high CCV_step2 threshold (H/xx) was chosen assuming that this represents the case where as much information as possible about the citation data is exploited, i.e., by using the two expanded citation indexes and a high threshold in Step 2. However, to allow the best match boomerang effect to perform at its best, the run with the highest AvgP score was chosen from this combination regardless of the DCV_step1 value.

Table 7.11 summarises the AvgP scores of the tested runs, as well as the results of the statistical tests. When measured by the str_inex quantification function no statistically

significant differences could be identified between the runs. For the other three quantification functions, the bag-of-words baseline showed significantly better performance than the best match boomerang effect and the polyrepresentation baseline. There was no significant difference between the best match boomerang effect and the polyrepresentation baseline with any of the quantification functions.

Figure 7.2 shows P-R curves of the three runs for all four quantification functions. Compared to Figure 7.1 it can be observed that the bag-of-words baseline remains at the top of all curves in Figure 7.2, although the improvements in the best match boomerang effect and polyrepresentation baseline mean that these are *on par* with the bag-of-words baseline at a few points. The best match boomerang effect now performs much more similar to the polyrepresentation baseline, and their curves are intertwined at several points in all quantification functions. The low precision found even at low recall levels for the best match boomerang effect in the official INEX 2002 results have now disappeared, and the best match boomerang effect achieves the same level of precision as the polyrepresentation baseline also at low levels of recall. The best match boomerang effect does, however, drop below the polyrepresentation baseline around 0.25 recall in all but the str_inex quantification function.

Figure 7.3 and Figure 7.4 show (nD)CG curves of the three runs. Two different sets of gain values were used corresponding roughly to the strict and generalized quantification functions used in the computation of AvgP values. Only the topical relevance dimension was taken into account though – the coverage dimension was ignored. The 0-1-2-3 combination of gain values awards the face value of the relevance assessment to each retrieved document, i.e., a document assessed as Highly relevant is given the score 3 in the computation etc. This corresponds most closely to the gen_whole quantification function (See Figure 6.9), except for the 2S and 1S assessments, which were excluded in the gen_whole quantification function. The 0-0-0-3 combination of gain values only gives award for retrieved documents with topical relevance degree 3. This corresponds to the str_whole quantification function.

Both (D)CG and n(D)CG curves are shown because the differences at low ranks are indistinguishable in the (D)CG curves. The n(D)CG curves facilitate interpretation of the results at lower ranks. The (nD)CG curves are particularly useful when the recall

base is sparse as in INEX, because they show the performance of the IR techniques in relation to the actual recall base.

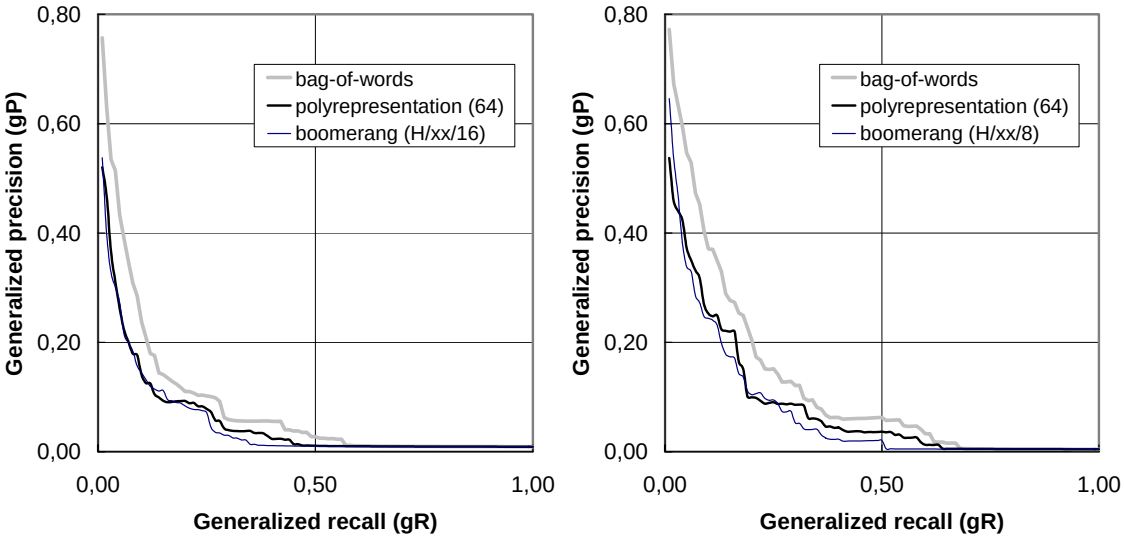
The (nD)CG curves in both Figure 7.3 and Figure 7.4 all show the same tendencies. The bag-of-words baseline is at the top of all runs closest to the ideal vector. There is no point on any of the curves where either the best match boomerang effect or the polyrepresentation baseline is on par with the bag-of-words baseline as observed on the P-R curves in Figure 7.2. The best match boomerang effect and the polyrepresentation baseline perform quite similarly and are intertwined at several points on the curves, especially with the 0-0-0-3 gain values. Initially, the best match boomerang effect lies slightly below the polyrepresentation baseline, except for the very first document in the 0-0-0-3 gain values. The best match boomerang effect obtains slightly better performance than the polyrepresentation baseline in the 20 to 40 rank interval, but its performance decreases slowly from rank 60 and upwards compared to the polyrepresentation baseline. An advantage of the n(D)CG curves (curves c and d in Figure 7.3 and Figure 7.4) is that the tendencies at rank 1 to 10 become much clearer compared to the (D)CG curves. It is interesting to note in the n(D)CG curves that all three runs actually manage to retrieve of fair proportion of the relevant documents in relation to what is theoretically possible given the recall base in INEX: the best match boomerang effect and polyrepresentation baseline retrieve from 30 to 60 percent of the recall base, and the bag-of-words baseline from 45 to 80 percent of the recall base.

**Table 7.11. Statistical results and AvgP scores for the runs used in the test of research question 1 (Summarised from Table 7.6 to Table 7.9). Values in bold face denote the best run in each column.**

| | gen_whole | str_whole | gen_inex | str_inex |
|---|---|---|---|---|
| Boomerang (H/xx) | 0.0508 | 0.0708 | 0.0392 | 0.0652 |
| Polyrepresentation baseline | 0.0554 | 0.0799 | 0.0419 | 0.0678 |
| Bag-of-words baseline | **0.0837** | **0.1193** | **0.0606** | **0.0830** |

DCV_step1 = 16 for the best match boomerang effect gen_whole and gen_inex quantification functions, and DCV_step1 = 8 for str_whole and str_inex. For the polyrepresentation baseline DCV_step1 = 64.

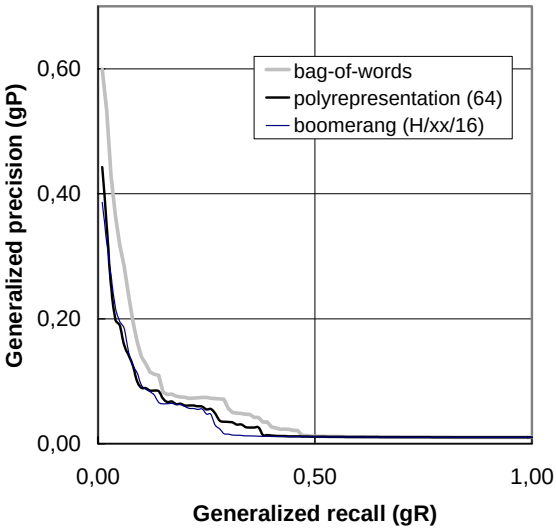
| Statistical significance (p = 0.05, gen_whole, str_whole and gen_inex): |
|---|
| bag-of-words > polyrepresentation, boomerang |



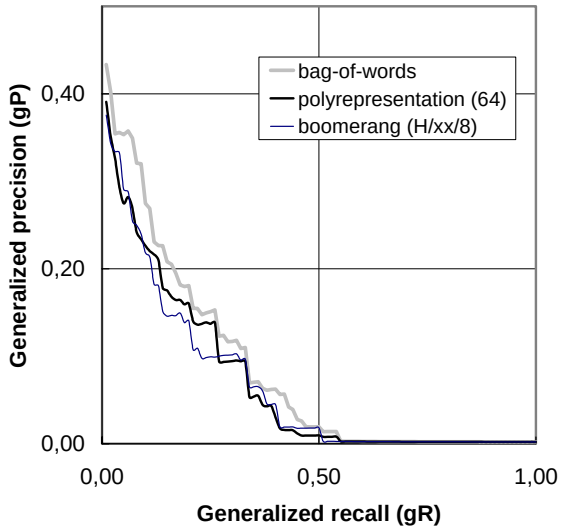

a) gen_whole




b) str_whole

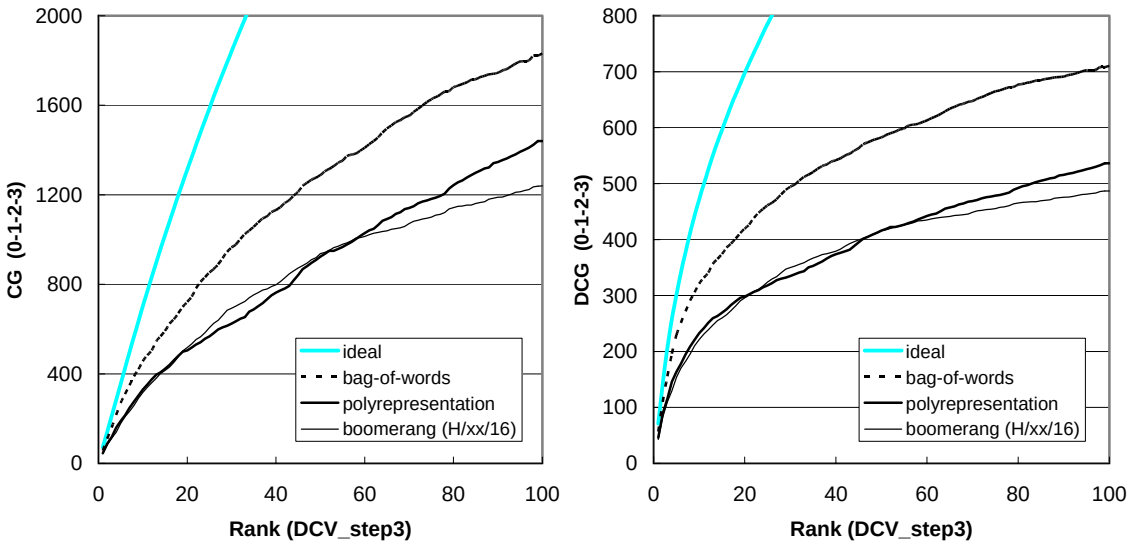



c) gen_inex



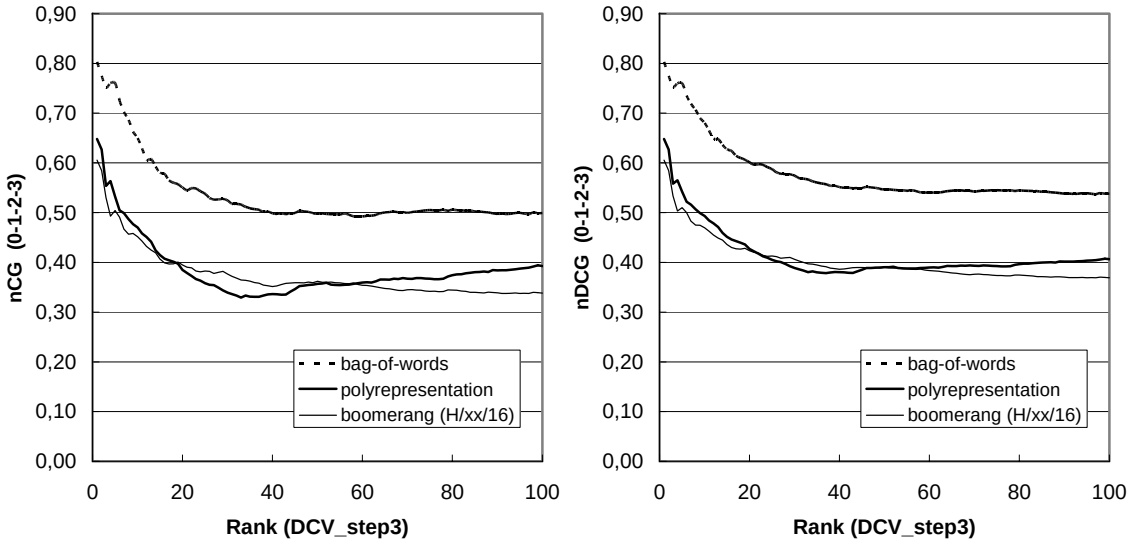

d) str_inex

**Figure 7.2.a-d. P-R curves of the best match boomerang effect and the baseline runs as tested in research question 1 for the four quantification functions. Note that the y-axis is fitted to each curve.**




a) CG (0-1-2-3)

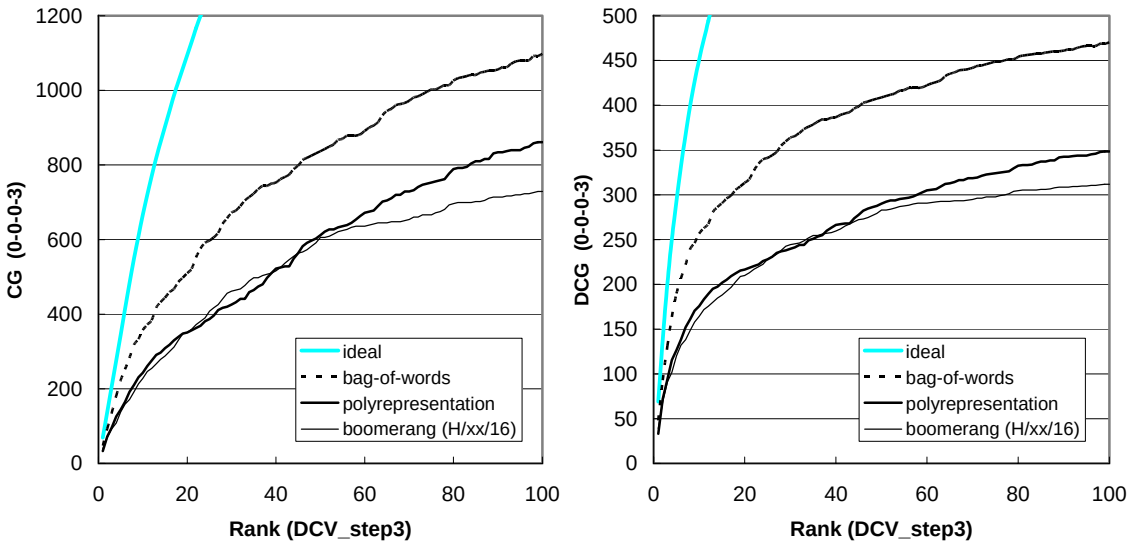
b) DCG (0-1-2-3)




c) nCG (0-1-2-3)

d) nDCG (0-1-2-3)

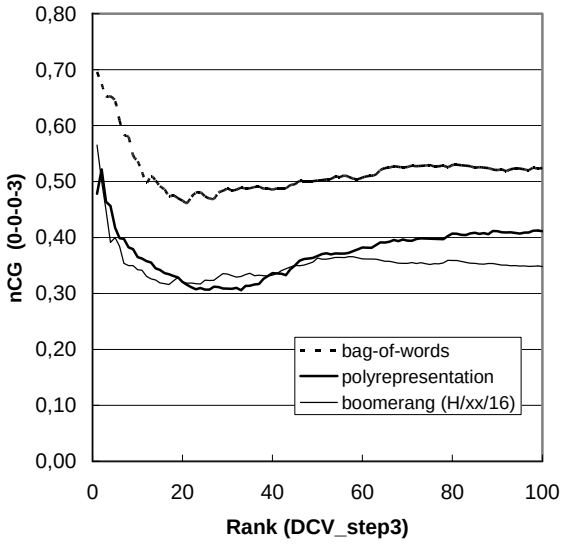
**Figure 7.3.a-d. (nD)CG curves of the best match boomerang effect and the baseline runs as investigated in research question 1. Gain values for the computations were 0-1-2-3, and the natural logarithm (2.718) was used as discount factor for the (n)DCG computations.**



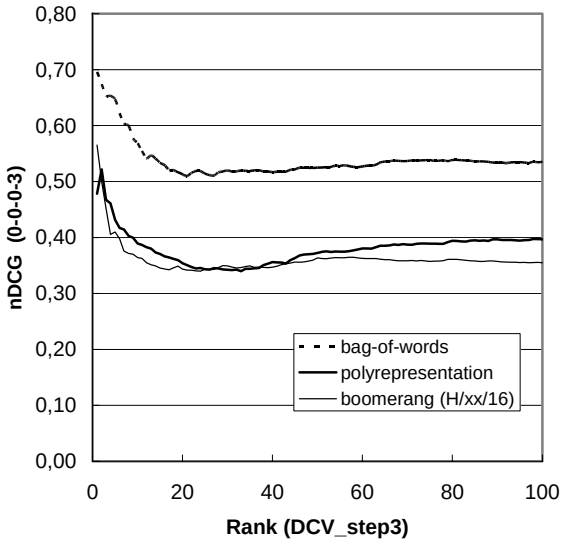

a) CG (0-0-0-3)




b) DCG (0-0-0-3)




c) nCG (0-0-0-3)




d) nDCG (0-0-0-3)

**Figure 7.4.a-d. (nD)CG curves of the best match boomerang effect and the baseline runs as investigated in research question 1. Gain values for the computations were 0-0-0-3, and the natural logarithm (2.718) was used as discount factor for the (n)DCG computations.**

#### 7.2.3.2 Research question 2: Are there significant differences in IR characteristics and performance between individual representations of the scientific full text documents?

This question investigates if there were significant differences in IR characteristics and performance between the ten individual representations of the scientific full text documents. Only the performance differences are examined here. Results relating to other IR characteristics have been presented in Section 7.2.1 above and these are discussed in Chapter 8 together with the performance differences presented here. The AvgP scores of the representations are listed in Table 7.10 above, and visualised in Figure 7.5. All ten runs were tested against each other. Table 7.12 summarises the results of the test. For the str_inex quantification function, the only significant difference was that the cited titles performed better than the introductions and conclusions. The remaining three quantification functions agree in the majority of cases. Therefore, only the significant differences where all three agree are reported in the following. The cited titles perform significantly better than most of the representations, except for the article titles, abstracts and identifiers. The abstracts and identifiers also perform better than the author keywords, conclusion and table captions, and the article titles better than the last two. In summary, there is a *top group* consisting of the cited titles, abstracts, identifiers and article titles with no significant differences among themselves, but with many significant differences to the worst performing runs. The latter consist of the table captions, conclusions and to a certain extent the author keywords and figure captions. The descriptors and introductions are rarely significantly different from any of the other runs.

**Table 7.12. Statistical results for the differences between the individual representations (research question 2). For the gen_whole, str_whole and gen_inex quantification functions only the differences where all three agree are reported.**

Statistical significance (p = 0.05, gen_whole, str_whole and gen_inex):

CTI > DE, INT, FGC

CTI, ABS, ID > AKW, CON, TBC

ATL > CON, TBC

Statistical significance (p = 0.05, str_inex):

CTI > INT, CON

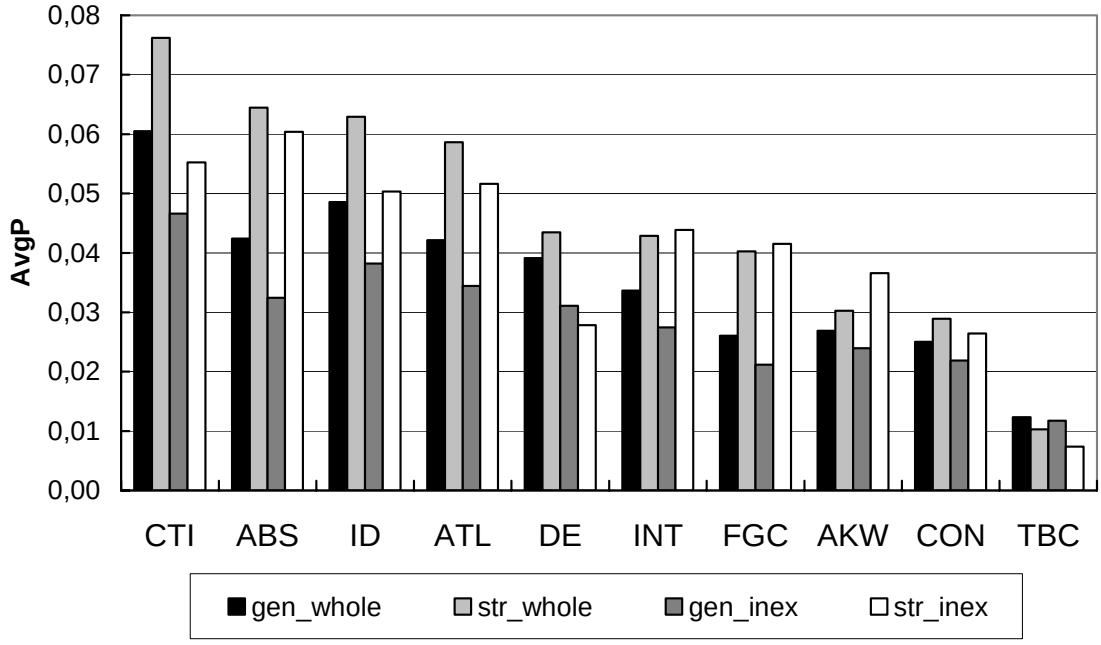



**Figure 7.5. Visualisation of the performance of the 10 individual representations tested in research question 2.**

#### 7.2.3.3 Research question 3: Does an increase of the number of source documents in the simple polyrepresentation baseline improve performance?

The question investigates if the performance of the polyrepresentation baseline can be improved by increasing the number of documents used in the calculation of the final weights in the polyrepresentation baseline. This number of documents is determined by the DCV_step1 threshold. An increased number of documents might result in better performance because of the increase in the statistical evidence from which the final rank is formed.

Research question 3 was investigated by testing the statistical significance of the differences between the runs over the whole range of DCV_step1 values for all four quantification functions (See Table 7.6 to Table 7.9 for the AvgP scores of the runs). A total of four Friedman tests were carried out. The results of the tests are reported in Table 7.13 for all four quantification functions.

No significant differences were identified for the str_inex function.

For the other three functions the tests reveal that the runs in the middle range around DCV_step1 = 64 are not significantly different from each other. The significant differences are mainly found between these middle range runs on one side, and the low (2, 4 and 8) and high (256, 512) ends of the range of DCV_step1 values on the other.

Overall, the test results show that the polyrepresentation baseline performs best at DCV_step1 values between 16 and 128, and that the runs at DCV_step1 = 256 or more perform significantly worse. Thereby an increase up to between 16 and 128 source documents improves performance in the polyrepresentation baseline. 256 source documents or more decreases performance.

**Table 7.13. Overview of the statistical differences tested in research question 3 between the runs with different DCV_step1 values. Results are reported for all four quantification functions.**

| Statistical significance (p = 0.05): | |
|---|---|
| gen_whole | 064, 256, 512 > 008, 004, 002<br>032 > 004, 002<br>128, 016 > 256, 512, 002 |
| str_whole | 064 > 256, 512, 008, 004, 002<br>032 > 512, 004, 002<br>128 > 004, 002<br>016, 256 > 002 |
| gen_inex | 064 > 256, 512, 008, 004, 002<br>032 > 512, 008, 004, 002<br>128, 016, 256 > 004, 002<br>512 > 002 |
| str_inex | no significant differences |

#### 7.2.3.4 Research question 4: Does performance improve by increasing the number of documents from which references are extracted in the best match boomerang effect?

Increasing the number of documents in Step 1 of the best match boomerang effect (i.e., increasing the DCV_step1 value) will increase the number of references that are extracted for the pools in Step 2, and will also increase the number of seed documents used in the weighted sum query in Step 3 (See Table 7.4). As discussed in Chapter 3, earlier studies have used very few intellectually selected seed documents. A significantly larger amount of seed documents might increase performance, either overall or at certain levels of recall. The question investigates if performance can be improved by increasing the number of source documents from which references are extracted in Step 1 of the best match boomerang effect. The effect over recall levels is also illustrated.

As described above in Section 7.2.2 the AvgP scores tend to peak at relative low values of DCV_step1, with lower AvgP scores for both low and high values of DCV_step1. The statistical significance of these differences was tested over the range of DCV_step1 values of each row in Table 7.6 to Table 7.9 (a total of 48 Friedman tests). All tests found at least one significant difference in each row, and pair wise comparisons between each of the runs in each row were subsequently carried out (a total of 1728 pair wise comparisons). An overview of these comparisons is given in Figure 7.2 in the form of a lower-left matrix for each quantification function. Each matrix summarises the number of times a significant difference was found between each of the DCV_step1 values. Since there are 12 rows in the tables, each cell in the matrix can represent from zero to 12 significant differences. Different tones of shading have been applied to the cells in Figure 7.6 to provide a better overview. A darker shade indicates that more significant differences were identified. A white cell indicates that no statistically significant difference was found.

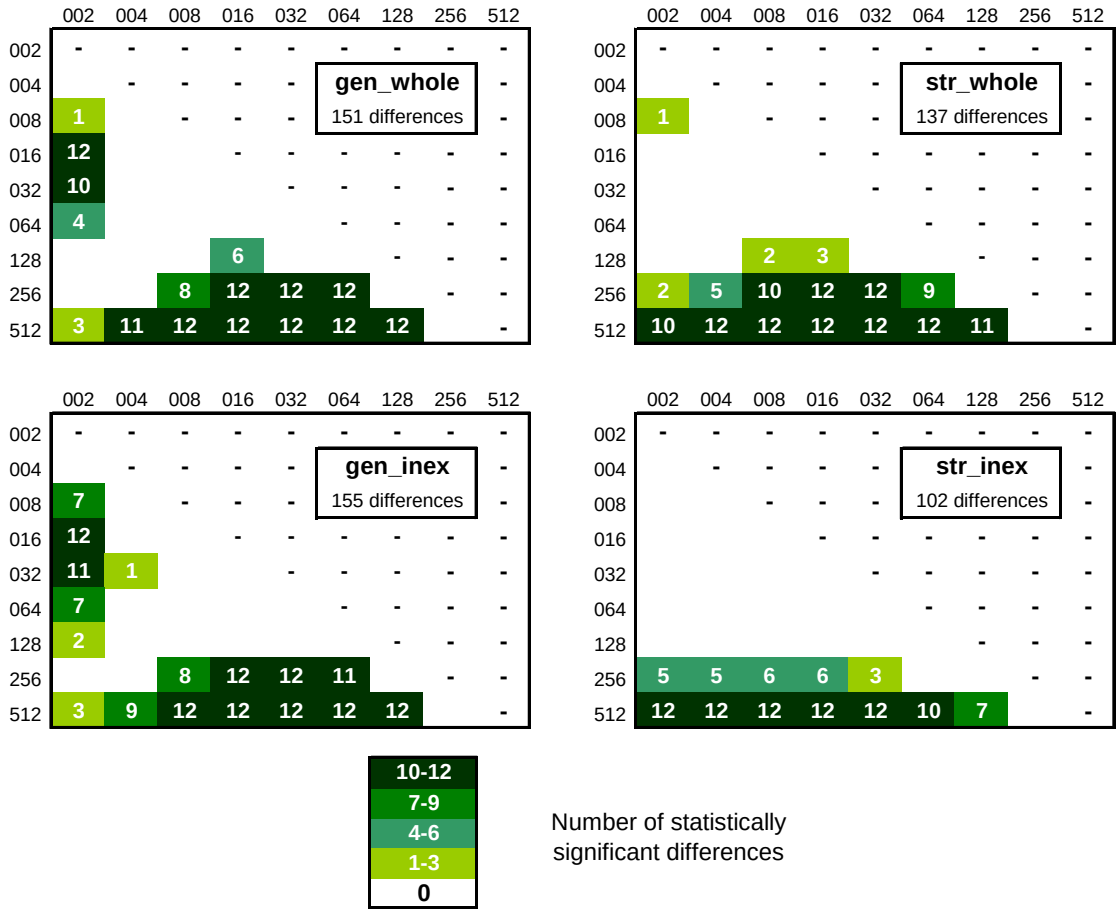


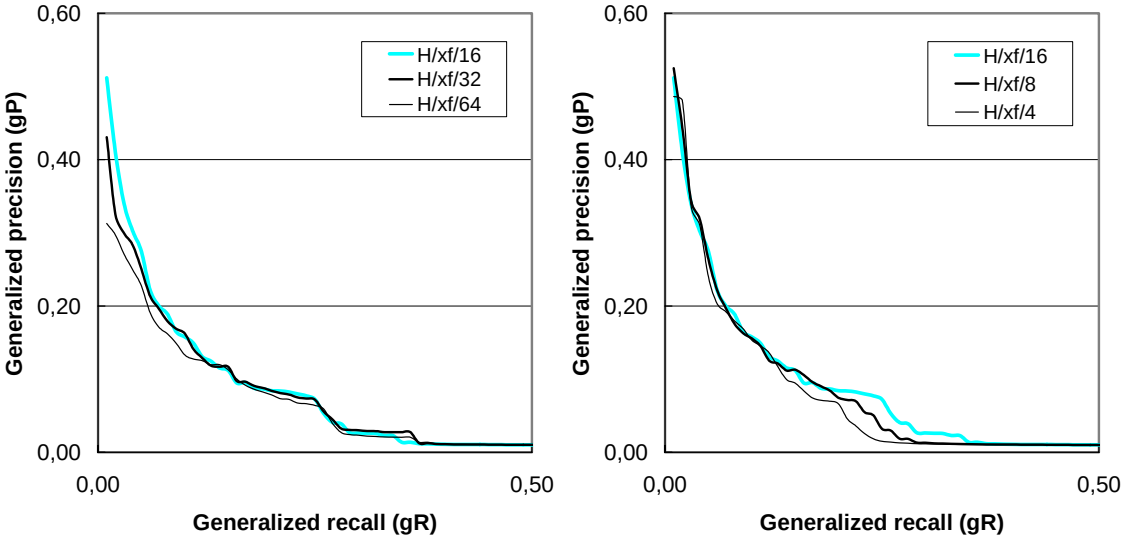

**Figure 7.6. Overview of the statistically significant differences over all ranges of DCV_step1 (p = 0.05). The matrices summarise all best match boomerang effect rows in Table 7.6 to Table 7.9. Dark shades between two runs indicate that more differences were significant between the two.**

For the two generalized quantification functions it may be observed that there were very few significant differences in the middle range of the DCV_step1 values: significant differences were mainly identified with the lowest value (002) and the highest values (512 and 256). For the strict quantification functions significant differences were predominantly identified with the highest DCV_step1 values (512, 256, and to a certain extent 128). As with the generalized functions very few differences could be identified in the middle range of the DCV_step1 values. Overall it may therefore be concluded that although the lowest value of DCV_step1 did result in significantly inferior performance (when measured with the generalized functions) there were not any significant differences when extracting references from 4 to at least 64 documents in Step 1 of the best match boomerang effect. Only the runs with the very highest values of

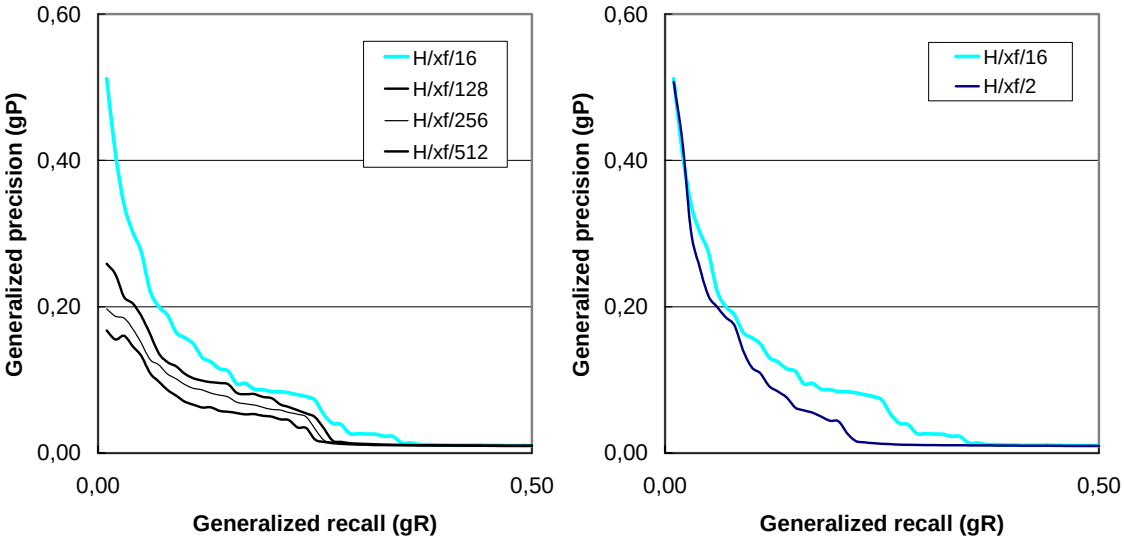
DCV_step1 showed significantly inferior performance. Note that because of the aggregation carried out in order to provide the overview in Figure 7.6, a considerable part of the underlying details has been reduced. Therefore, in the cases where less than 12 differences were identified, it is not immediately possible to observe which of the other variables apart from DCV_step1, which did not result in significant differences. Consequently, it is not possible to study interactions between the variables in Figure 7.6. Such interactions can, however, be studied with the data from research question 5 to 7.

In order to illustrate the findings from Figure 7.6 in detail, the runs over the whole range of DCV_step1 values in a single row from Table 7.6 are displayed as P-R curves in Figure 7.7. The particular combination of CCV_step2 = High and the citation indexes (xf) was chosen deliberately because it had a relatively high number of significant differences between the runs. The run with the best absolute AvgP value (DCV_step1 = 16) is shown in each of the four P-R curves in the figure, together with the runs that are either significantly different (curves a and b) or not significantly different (curves c and d) from this run. In each case the runs have been divided into those that have a higher and lower DCV_step1 value than 16 respectively in order to study differences between them. It may be observed that none of the runs in Figure 7.7 achieve very high levels of recall – the highest level is achieved by DCV_step1 = 64, which reaches 0.36 recall at its maximum. For the runs that are significantly different from the best one *and* have higher value of DCV_step1 (i.e., 512, 256 and 128, curve c) it can be seen that the reason for these differences is that these runs achieve markedly lower precision, especially at low recall levels. The run with lower DCV_step1, that is significantly different from the best one (i.e., DCV_step1 = 2, curve d), follows the opposite pattern: it starts out with the same high level of precision at the lowest recall level, but quickly drops below this. The same pattern may be observed with the runs that are not significantly different from the best one: Although closely intertwined with this, the runs at higher values of DCV_step1 (32 and 64, curve a) achieve slightly lower precision at low recall levels and slightly higher at high recall levels. The runs with lower values of DCV_step1 have approximately the same level of precision as the best one until their precision drop at higher levels of recall (DCV_step1 = 4 and 8, curve b). The example illustrates that high precision can be achieved initially at low values of DCV_step1, but that a higher value of DCV_step1 is needed in order to maintain this at

high recall levels. At the highest values of DCV_step1 precision suffers even at low recall levels.




a) 16, 32 and 64 (no sign. difference) b) 16, 8 and 4 (no sign. difference)




c) 16 > 512, 256, 128 (p = 0.05) d) 16 > 2 (p = 0.05)

**Figure 7.7.a-d. P-R curves of all DCV_step1 values of a particular combination of CCV_step2 (High) and citation indexes (xf) for the gen_whole quantification function. The best absolute AvgP run (DCV_step1 = 16) is shown together with the runs that were not significantly different from it (a and b), as well the runs that were (c and d). Note that only up to 0.50 recall is shown.**

#### 7.2.3.5 Research question 5: Can better performance be obtained by reducing the number of seed documents in the best match boomerang effect to the seed documents with the highest weights?

The question investigates if better performance can be obtained by reducing the number of seed documents in the forward chaining in Step 3 of the best match boomerang effect. The reduction is achieved by using the Medium and Low levels of the CCV_step2 threshold instead of the High level (See Section 6.3.2 above). As described in Section 7.2.2 there was a tendency that the absolute AvgP values increased over the three values of CCV_step1, although less so at the highest values of DCV_step1.

This tendency was examined by testing for the statistical significance between the runs made with Low, Medium and High values of the CCV_step1 threshold for every combination of DCV_step1 and the citation indexes (Table 7.6 to Table 7.9). A total of 144 Friedman tests were carried out. In the cases where the test indicated that at least one of the runs was significantly different from one of the other runs, pair wise comparisons were made in order to identify which ones which were different. There were three possible types of differences in each of the 144 tests:

A) A difference exists between the Low and the Medium run,
B) A difference exists between the Low and the High run, and
C) A difference exists between the Medium and the High run.

Table 7.14 provides an overview of the results of the statistical tests using this notation. Overall, it can be seen that there were relatively few and scattered significant differences. Of the 324 possible differences in all quantification functions only 35 were significant. As expected, the most frequent one was type B, which occurred 22 times. Type A occurred 12 times, and type C only 1 time. In seven cases type A and type B occurred together. For the gen_whole, str_whole and gen_inex quantification functions the majority of the significant differences occurred for the xf and xx combination of the citation indexes. The type A differences all occurred at DCV_step1 = 64 or lower, whereas the type B differences tended to be either at the two extremes of the DCV_step1 range or in the middle of it. On the whole, the tests reveal that the apparent differences between the three levels of the CCV_step2 threshold described in Section 7.2.2 are only statistically significant in a limited number of cases. In summary, no general improvements in performance could be demonstrated when reducing the seed documents to those with the highest weights.

**Table 7.14. Overview of the statistical differences (p = 0.05) tested in research question 5 between the three values of CCV_step2 (Low, Medium, High). Differences are reported for all combinations of citation indexes (CI), and over the whole range of DCV_step1 values for all four quantification functions. A significant difference between Low and Medium is indicated an "A", between Low and High by a "B", and between Medium and High be a "C".**

| Quantification function | CI | 2 | 4 | 8 | 16 | 32 | 64 | 128 | 256 | 512 |
|---|---|---|---|---|---|---|---|---|---|---|
| gen_whole | ff | --- | --- | --- | -B- | --- | --- | --- | --- | -B- |
| gen_whole | fx | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| gen_whole | xf | AB- | --- | --- | -B- | AB- | --- | --- | --- | --- |
| gen_whole | xx | AB- | --- | --- | --- | --- | -B- | --- | --- | --- |
| str_whole | ff | --- | --- | --- | --- | --- | --- | --- | --- | -B- |
| str_whole | fx | --- | --- | --- | --- | --- | --- | --- | --- | -B- |
| str_whole | xf | -B- | A-- | --- | --- | AB- | --- | --- | --- | --- |
| str_whole | xx | AB- | A-- | --- | --- | --- | A-- | --- | --- | --- |
| gen_inex | ff | --- | --- | --- | --- | --- | --- | --- | --- | -B- |
| gen_inex | fx | --- | --- | --- | --- | --- | --- | --- | -B- | --- |
| gen_inex | xf | AB- | --- | --- | -B- | --- | --- | --- | --- | --- |
| gen_inex | xx | --- | --- | --- | -B- | AB- | -B- | --C | --- | --- |
| str_inex | ff | --- | --- | --- | --- | --- | --- | --- | --- | -B- |
| str_inex | fx | --- | --- | --- | --- | --- | --- | --- | -B- | -B- |
| str_inex | xf | --- | --- | --- | --- | --- | A-- | --- | --- | --- |
| str_inex | xx | --- | --- | --- | --- | --- | A-- | --- | --- | --- |

#### 7.2.3.6 Research question 6: Can better performance be obtained by using an expanded citation index as basis for the weighting and selection of citations in the boomerang effect compared to a flat citation index?

In the expanded citation index each citation occurs as many times as it is mentioned in the full text of the document, not just once per document as in the flat, conventional citation indexes (See Section 6.2.4). Research question 6 investigates if better performance can be achieved by using the expanded citation index for the extraction and weighting of citations in Step 2 of the best match boomerang effect. This might

result in better performance because the expanded citation index provides more statistical evidence for each citation in the matrix in Step 2.

Research question 6 was investigated by testing for the statistical significance between the runs where the flat citation index was used for the extraction and weighting of citations in Step 2 against the runs where the expanded citation index was used. The tests were carried out for every possible combination of DCV_step1 and CCV_step2 and the citation indexes. Thus two types of pairs were available for the tests: ff runs versus xf runs, and fx runs versus xx runs (indicated in Table 7.15 as ff-xf, and fx-xx). In both pairs the citation index used in Step 2 vary (the underlined f or x) and the citation index used in Step 3 is fixed. A total of 216 Friedman tests[66] were performed (54 for each quantification function). Table 7.15 provides an overview of the tests. If there was a significant difference between two runs the best performing run is indicated in the table.

Overall it may be observed that there were statistically significant differences in slightly more than half the cases (116/216). In these 116 cases the best performing runs occurred when a flat citation index (ff or fx) was used for the extraction and weighting of citations in Step 2 of the best match boomerang effect. That is, the expanded citation index did *not* perform significantly better than the flat in a single case when there was a statistically significant difference.

The majority of the significant differences were to be found at values of DCV_step1 of 32 and above across all four quantification functions. The gen_whole, str_whole, and gen_inex functions displayed very similar patterns: at the 32-512 DCV_step1 range all differences were significant. The remaining ones were more scattered, but concentrated at DCV_step1 = 16, and at the High value of CCV_step2. In total, 62 % (100/162) of the cases in the three functions were significant, and there was little difference between the two types of pairs (ff versus xf, and fx versus xx). All significant differences in the str_inex quantification function occurred in the 32-512 DCV_step1 range, but they were more scattered than in the other three functions and fewer: only 30 % (16/54) of the cases showed significant differences.

[66] Which in this case reduces to a sign test, because there are only two runs in each test.

On the whole, the statistical tests of research question 6 revealed that the majority of the runs in the 32-512 DCV_step1 range did not benefit from an expanded citation index for the extraction and weighting of citations in Step 2 of the best match boomerang effect. For the DCV_step1 values below 32 there were few significant differences between using an expanded or a flat citation index in Step 2. In summary, no performance gains could be demonstrated by using an expanded citation index as basis for the weighting and selection of citations in the boomerang effect compared to a flat citation index.

**Table 7.15. Overview of the statistical differences tested in research question 6 between the runs where a flat citation index was used against the runs where an expanded citation index was used for the extraction and weighting of citations in Step 2. Results are reported for the different combinations of these (ff versus xf, and fx versus xx), and over CCV_step2 values for all four quantification functions. In case of a significant difference (at p = 0.05) between two runs the run which performed better is indicated in the table.**

| Quantification function | Runs | CCV | 2 | 4 | 8 | 16 | 32 | 64 | 128 | 256 | 512 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| gen_whole | ff-xf | L | - | - | - | ff | ff | ff | ff | ff | ff |
| gen_whole | ff-xf | M | - | - | - | - | ff | ff | ff | ff | ff |
| gen_whole | ff-xf | H | - | - | ff | ff | ff | ff | ff | ff | ff |
| gen_whole | fx-xx | L | - | - | - | fx | fx | fx | fx | fx | fx |
| gen_whole | fx-xx | M | - | - | - | - | fx | fx | fx | fx | fx |
| gen_whole | fx-xx | H | - | - | fx | fx | fx | fx | fx | fx | fx |
| str_whole | ff-xf | L | - | - | - | - | ff | ff | ff | ff | ff |
| str_whole | ff-xf | M | - | - | - | - | ff | ff | ff | ff | ff |
| str_whole | ff-xf | H | - | - | - | ff | ff | ff | ff | ff | ff |
| str_whole | fx-xx | L | - | - | - | - | fx | fx | fx | fx | fx |
| str_whole | fx-xx | M | - | - | - | - | fx | fx | fx | fx | fx |
| str_whole | fx-xx | H | - | - | - | - | fx | fx | fx | fx | fx |
| gen_inex | ff-xf | L | - | - | - | ff | ff | ff | ff | ff | ff |
| gen_inex | ff-xf | M | - | - | - | - | ff | ff | ff | ff | ff |
| gen_inex | ff-xf | H | - | ff | - | - | ff | ff | ff | ff | ff |
| gen_inex | fx-xx | L | - | - | - | - | fx | fx | fx | fx | fx |
| gen_inex | fx-xx | M | - | - | - | - | fx | fx | fx | fx | fx |
| gen_inex | fx-xx | H | - | - | - | fx | fx | fx | fx | fx | fx |
| str_inex | ff-xf | L | - | - | - | - | ff | ff | ff | ff | ff |
| str_inex | ff-xf | M | - | - | - | - | - | ff | - | ff | - |
| str_inex | ff-xf | H | - | - | - | - | ff | - | - | ff | - |
| str_inex | fx-xx | L | - | - | - | - | - | - | fx | fx | - |
| str_inex | fx-xx | M | - | - | - | - | - | fx | fx | fx | - |
| str_inex | fx-xx | H | - | - | - | - | - | - | fx | fx | - |

#### 7.2.3.7 Research question 7: Can better performance be obtained by running the citation queries against an expanded citation index of the boomerang effect compared to a flat citation index?

Research question 7 investigates if better performance can be achieved by using the expanded citation index as the index against which the weighted sum citation queries are run in Step 3 of the best match boomerang effect. The use of the expanded citation index might result in better performance because it provides more statistical evidence about each reference in the documents as discussed in Section 6.2.4 above.

Research question 7 was tested as research question 6, but by swapping the variable that was fixed. The statistical significance of the difference between the runs, where the flat citation index was used in Step 3, was tested against the runs where the expanded citation index was used in Step 3. The tests were carried out for every possible combination of DCV_step1 and CCV_step2 and the citation indexes. Therefore two types of pairs were available for the tests: ff runs versus fx runs, and xf runs versus xx runs (indicated in Table 7.16 as ff-fx, and xf-xx). In both pairs the citation index used in Step 3 varies (the underlined f or x) and the citation index used in Step 2 is fixed. A total of 216 Friedman tests[67] were performed (54 for each quantification function). Table 7.16 provides an overview of the tests. If there was a significant difference between two runs the best performing run is indicated in the table.

Overall it may be observed that there were statistically significant differences in 39 % (85/216) of the cases. In these 85 cases the best performing runs occurred when a flat citation index (ff or xf) was used in Step 3. That is, the expanded citation index did *not* perform significantly better than the flat in a single case when a statistically significant difference could be identified.

The majority of the significant differences were to be found at values of DCV_step1 of 32 and above across all four quantification functions. The generalized functions displayed very similar patterns: at the 32-512 DCV_step1 range, almost all differences were significant. The few remaining ones were more scattered, and occurred at the High value of CCV_step2. In total 54 % (58/108) of the cases in the generalized functions were significant, and there was little difference between the two types of pairs (ff versus

[67] Which in this case reduces to a sign test, because there are only two runs in each test.

fx, and xf versus xx). For both of the strict functions the majority of the significant differences occurred in the 32-512 DCV_step1 range, but they were more scattered than in the other two functions and also fewer: only 25 % (27/108) of the cases showed significant differences. Contrary to the generalized functions, the strict functions had very few differences at DCV_step1 = 512. With the strict functions there were differences between the two types of pairs: at the xf-xx pair the significant differences tended to be concentrated at 128 and 256 values of DCV_step1.

On the whole, the statistical tests of research question 7 revealed that the majority of the runs in the 32-512 DCV_step1 range did not benefit from an expanded citation index, as the index against which the weighted citation queries were run in Step 3 of the best match boomerang effect. This was especially apparent with the generalized functions. For the lower DCV_step1 values there were few significant differences.

**Table 7.16. Overview of the statistical differences tested in research question 7 between the runs where a flat citation index was used against the runs where an expanded citation index was used for the extraction and weighting of citations in Step 3. Results are reported for the different combinations of these (ff versus fx, and xf versus xx), and over CCV_step2 values for all four quantification functions. In case of a significant difference (at p = 0.05) between two runs the run that performed better is indicated in the table.**

| Quantification function | Test | CCV | 2 | 4 | 8 | 16 | 32 | 64 | 128 | 256 | 512 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| gen_whole | ff-fx | L | - | - | - | - | ff | ff | ff | ff | - |
| gen_whole | ff-fx | M | - | - | - | - | ff | ff | ff | ff | ff |
| gen_whole | ff-fx | H | - | - | ff | ff | ff | ff | ff | ff | ff |
| gen_whole | xf-xx | L | - | - | - | - | xf | - | xf | xf | xf |
| gen_whole | xf-xx | M | - | - | - | - | xf | xf | xf | xf | xf |
| gen_whole | xf-xx | H | - | - | xf | - | xf | xf | xf | xf | xf |
| str_whole | ff-fx | L | - | - | - | - | - | - | ff | - | - |
| str_whole | ff-fx | M | - | - | - | - | ff | ff | ff | - | - |
| str_whole | ff-fx | H | - | - | - | - | ff | ff | - | - | - |
| str_whole | xf-xx | L | - | - | - | - | - | - | xf | xf | - |
| str_whole | xf-xx | M | - | - | - | - | xf | - | xf | xf | - |
| str_whole | xf-xx | H | - | - | - | xf | - | - | xf | xf | - |
| gen_inex | ff-fx | L | - | - | - | - | - | ff | ff | ff | ff |
| gen_inex | ff-fx | M | - | - | - | - | - | ff | ff | ff | ff |
| gen_inex | ff-fx | H | - | - | - | ff | ff | ff | - | ff | ff |
| gen_inex | xf-xx | L | - | - | - | - | - | - | xf | xf | xf |
| gen_inex | xf-xx | M | - | - | - | - | xf | xf | xf | xf | xf |
| gen_inex | xf-xx | H | - | - | - | xf | xf | xf | xf | xf | xf |
| str_inex | ff-fx | L | - | - | - | - | - | - | - | ff | ff |
| str_inex | ff-fx | M | - | - | - | - | - | ff | ff | - | - |
| str_inex | ff-fx | H | - | - | ff | - | - | ff | - | ff | ff |
| str_inex | xf-xx | L | - | - | - | - | - | - | - | xf | - |
| str_inex | xf-xx | M | - | - | - | - | - | - | xf | xf | - |
| str_inex | xf-xx | H | - | - | - | - | - | - | xf | xf | - |

## 7.3 Summary statements

This Chapter has described the results of the preliminary INEX submission, and the results of the main experiment. Descriptive data on the behaviour of the best match boomerang effect and the baselines were presented, and a number of overall trends were identified in the result. The results of each research questions were dealt with one by one. The discussion of the results is placed in Chapter 8 because there seems to be interactions between the variables and the research questions.

# 8 Discussion

In this chapter the analytical, methodological and empirical results from chapters 2 to 6 are discussed. Some of the research questions tested in Chapter 7 are discussed together here as they either are very similar in scope, or appear to interact with each other. The chapter ends with summary statements in Section 8.4.

The cognitive viewpoint in general and the theory of polyrepresentation in particular seems a good platform from which to do IR retrieval and seeking experiments because it is very holistic and amalgamates theories and empirical research results from both the system-driven and the user-oriented traditions. The idea of a polyrepresentation continuum is proposed in Chapter 3 as an extension to the theory of polyrepresentation. The polyrepresentation continuum has a structured and an un-structured pole emphasising the range of possible polyrepresentative opportunities along the continuum.

## 8.1 Analytical results

As shown by the reviewed studies in Chapter 4, as well as by the pre-experiment and the main experiment, references and citations can be used to retrieve relevant documents and often with good results. Chapter 4's analysis of research on citer motivations and citation behaviour did not result in a clear-cut explanation, or theoretical justification for *why* references and citations might be useful in IR. It rather identified a number of different positions that provide a range of reasons that may be helpful in explaining the achieved sound empirical results. The most unifying of these explanations is Small's conception of references as being *concept symbols* that stand for an idea that is being used in the course of an argument (Small, 1978). Although this conception does not explain why particular authors select particular references to symbolise particular concepts it does go some of the way towards bridging the normative and social constructivist positions identified in Chapter 3.

A number of search strategies as well as different applications in both laboratory and operational settings were described in Chapter 4 which gave examples of *how*

references and citations have been exploited in IR. The main part of these exploit *forward chaining,* a feature unique to citation indexes where unknown documents that refer to one or a few known seed documents are retrieved. As shown by the studies reviewed in Section 4.4, high precision can often be achieved with forward chaining. However, the studies also show that there is large variation in the performance across topics with this strategy (e.g., McCain, 1989). The main reason behind this variation is that the selection of suitable seed documents is crucial to the performance of forward chaining. As put by Pao: "Relevant retrieval from citation searching depends solely on good seed documents. Without them, citation searching simply cannot perform well." (1993, p. 108). The most important challenge in the use of references and citations in IR is thus to find improved methods to express the information need in terms of citations.

The boomerang effect proposed in Chapter 5 is a response to this challenge. The main difference between the earlier uses of references and citations in IR and the boomerang effect is that the latter only requires the user to specify a natural language query as in conventional IR systems. This is then automatically translated into seed documents and used in a forward chaining to retrieve documents that refer to these seed documents. Thereby the need for the user to select seed documents intellectually, and the limitations of this choice have been removed. One advantage of this is that many more seed documents can be used to represent the information need. Inspired by the theory of polyrepresentation (Ingwersen, 1996) the intention behind the boomerang effect is to reduce the extra uncertainty introduced by the automatic translation. This is done by emphasising those citations that occur in the overlaps between the documents identified by different cognitive and functional representations.

Two versions of the boomerang effect are presented in the dissertation: a Boolean version located towards the structured end of the polyrepresentation continuum, and a best match version at the unstructured end (See Figure 3.3). The Boolean version was tested in a pre-experiment in order to gain experiences that could be used in designing the main experiment. The pre-experiment demonstrated that it was possible to retrieve relevant documents through the network of references and citations with a Boolean version of the boomerang effect. This was achieved without the specification of seed documents in advance, and the automatic translation of the information needs into citation queries was thus shown to function. Further, the results of the pre-experiment

indicate that the concentration of relevant documents were larger in the top of the overlap structure generated by the Boolean boomerang effect.

The best match boomerang effect was tested in the main experiment. The results of this are discussed in Section 8.3 below.

## 8.2 Methodological results

Carrying out the main experiment with the INEX test collection had a number of advantages.

*Firstly*, without the INEX initiative it would not have been possible to carry out such extensive tests as done in the main experiment. Scientific documents in full text and formatted in such a way that functional representations can be extracted without great difficulty are very hard to get hold of in large numbers. This is the case even in the present day where most major scientific journal publishers produce their journals in exactly such formats (e.g., SGML or XML). INEX provides a 402 MB[68] corpus consisting of all 12,107 items from 18 magazines and journals over a 7-year period, in full text and formatted in XML. When measured by number the of documents this is by no means a large test corpus, e.g., compared to the size of TREC which has used 100,000s of documents. However, it is a unique characteristic of INEX that a realistic corpus of scientific and technical full text documents in XML is available for IR purposes.

*Second*, a number of topics and corresponding graded relevance assessments relating to the corpus were created in a collaborative effort by domain experts from the participating organisations. This exceeds by far the number of topics that realistically could have been created and assessed within the present dissertation work.

*Third*, the INEX organisers wisely decided to use *graded* relevance assessments in both relevance dimensions. This facilitates the computation of performance measures, which incorporate much of the recent research into the development of novel and more flexible

[68] 494 MB including the XML mark-up.

measures in response to the criticism against older and more rigid ones. This decision also makes it possible to use the INEX test collection for other purposes than those originally intended. For instance, although not intended for research with IR techniques that retrieve whole documents (see the discussion below), the construction of two new quantification functions which are better fitted to this type of retrieval, was possible without problems due to the availability of the graded relevance assessments.

*Fourth*, although both the launch and completion of INEX 2002 were rather late in relation to the dissertation work, the effective infrastructure provided optimal support for the main experiment, e.g. by providing a number of highly efficient tools that facilitated rapid computation of the performance measures.

There are two main *disadvantages* of using INEX in the main experiment: Firstly, the main purpose of INEX is to investigate the retrieval of *components* of documents, rather than whole documents as retrieved by all the IR techniques tested in the main experiment. Although retrieval of whole documents is possible with the CO topics, more emphasis was clearly put on the CAS topics as discussed in Section 6.1 above. The relatively low level of average precision of the CO topics, compared to the CAS topics (See Section 7.1 above), can be seen as a consequence of this. The optimistic relevance propagation serves to alleviate this to a certain extent because the recall base is expanded considerably, but the overall level of performance in INEX is still rather low for the CO topics. Consequently, there is a risk that differences between IR techniques are harder to detect than they would have been if more relevant documents had been identified on average for each topic, even when the significance level of the test is set at 0.05, as in the analysis of the main experiment. The Friedman test chosen for the tests of statistical significance is based on the *ranks* between IR techniques, and thus the absolute level of performance should not have any influence on the results of the tests[69]. Nevertheless, if the CO topics had been developed with the aim of identifying a greater proportion of relevant documents, larger relative differences between IR techniques might have been observable. The other objective one might have about using the INEX test collection for the main experiment is that it is limited to a particular area of research and technical development: the computer science area with a

[69] Multiplying all the data used for the tests by 10 would, for instance, not change the outcome of the test.

slant towards the engineering aspects. This makes it difficult to generalise the results beyond the subject areas covered by the collection, because differences in writing styles or citation behaviour might influence the results significantly. Ideally the main experiment should have been carried out on several different test collections. This has been practised in much of the laboratory research previously done in IR (See, e.g., the dissertation work by Ruthven (2001)). However, as discussed above, the INEX test collection is the first of its kind to contain full text scientific documents, including the references which are needed for the boomerang effect. Therefore, it is the only available alternative with which the full range of research questions could have been investigated. Some of the older test collections, which contain citation data, could have been used to test some of the questions. Research question 1, 2, 3, 4 and 5 could have been investigated, at least partly, using for instance, the Cystic Fibrosis test collection (Shaw, Wood and Tibbo, 1991). These older test collections do, however, not contain the full text, and they are relatively small both with regard to the number of documents and the total size (measured in MB).

The purpose of the gen_whole and str_whole quantification functions is to facilitate the evaluation of IR techniques that retrieve *whole* documents rather than components. The str_inex quantification function is particularly inappropriate for this purpose due to the relevance propagation rules as discussed in Section 6.4.1 above. The composition of the gen_whole and str_whole functions can be considered a success because the statistical significance tests based on the str_inex function showed less ability to differentiate between IR techniques on average. This tendency is obvious across all the results of each of the seven research questions (See Chapter 7). This is in contrast to the experiences of Järvelin and Kekäläinen (2000) and Kekäläinen and Järvelin (2002b) where the equivalent of the strict functions is better at distinguishing between IR techniques. The lesser ability of the str_inex quantification function to differentiate between IR techniques can therefore be seen as a confirmation of the appropriateness of replacing it with the str_whole function. A further indication that the str_inex function may yield unstable results is that the P-R curves based on it are much rougher and with sudden jumps along the curves, compared to the smoother curves produces by the other three functions (See Figure 7.1 and Figure 7.2). The differences between the results yielded by the gen_whole and the gen_inex quantification functions are not nearly as great. This resemblance is mainly due to the fact that they both assign relevance scores to most of the possible combinations of the two relevance dimensions, albeit with

slightly different values. In the following discussion the results obtained with the gen_whole, gen_inex and str_whole quantification functions will be given priority, as it is believed that the str_inex quantification function is not well suited for IR techniques that retrieve whole documents.

## 8.3 Empirical results

### *8.3.1 The official INEX 2002 runs and research question 1*

The low ranks of the boomerang run and the polyrepresentation baseline in the official INEX 2002 were particularly discouraging because each participant was allowed to submit up to three runs. The rank positions of the two runs (32-37 depending on quantification function) among the other 49 CO submissions were therefore in the worst performing third of the runs. The improvement in performance from the official INEX 2002 runs to the main experiment was, however, quite marked. The implementation of the DCV_step1 and CCV_step2 variables as well as the flat and expanded citation indexes were made in order to discover how these variables affect the performance of the best match boomerang effect and the polyrepresentation baseline.

With the experience gained from the main experiment, a failure analysis can be carried out of why the best match boomerang effect and the polyrepresentation baseline did not perform well in the official INEX 2002 results. The main factor that was changed, compared to the three official INEX 2002 submissions, was the topic elements included in the queries: The main experiment runs were made with the relatively short Title and Keywords elements, whereas the INEX 2002 submissions also included the longer Description element. The Title and Keyword elements contain few, selected terms that are deemed central to the information need expressed in the topic, while the Description element consists of a longer natural language sentence. Because no structure was imposed on the queries, any document containing just one of the query keys (excluding stop words) could be included in documents returned in Step 1. Thereby the source documents used for the polyrepresentation baseline and the best match boomerang effect would be of lower quality overall, resulting in lower performance. The polyrepresentation baseline and the best match boomerang effect are particularly sensitive to this because documents occurring in the overlaps are emphasised in both,

and this is probably the main reason for their lower performance in the official INEX 2002 results.

Note that there were no noticeable difference for the bag-of-words baseline between the official INEX 2002 submissions (Table 7.1) and the main experiment (Table 7.6 to Table 7.9). This indicates that the polyrepresentation baseline and the best match boomerang effect are very sensitive to the quality of the initial input. The bag-of-words baseline was probably less affected by the inclusion of the Descriptions because this scenario is much closer to what the weighting and ranking mechanism in InQuery is designed for. A second factor that seems to have decreased performance of the boomerang run in INEX 2002 is that the value of DCV_step1 was set far too high at 500 documents. As can be seen from Table 7.6 to Table 7.9, better performance could have been achieved, e.g., in the 8 to 32 DCV_step1 range. The relatively high levels of precision found in the top of the OLs in the pre-experiment were not found in the INEX 2002 boomerang run. Rather, the boomerang run displayed markedly lower precision at low levels of recall even compared to the polyrepresentation baseline (Figure 7.1), although it did slightly better than the polyrepresentation baseline at higher recall levels.

Research question 1 investigated whether the best match boomerang effect can achieve a similar level of performance as that of the bag-of-words baseline and the polyrepresentation baseline. As shown by the statistical tests, and as indicated by the absolute AvgP scores, the P-R curves, and the (nD)CG curves, the bag-of-words baseline performed significantly better than both the best match boomerang effect and the polyrepresentation baseline. Thus the improvements in performance compared to the official INEX 2002 results were not great enough to match the level of performance of the bag-of-words baseline. At the same time there is no significant statistical difference between the best match boomerang effect and the polyrepresentation baseline. This was also obvious from both the P-R curves (Figure 7.1) and the (nD)CG curves (Figure 7.2 and Figure 7.3) where the two runs are intertwined. The (nD)CG curves reveal that although the best match boomerang effect did marginally better than the polyrepresentation baseline in the 20 to 40 rank interval, it slowly decreased in performance from rank 60 and upwards. The reason for the superiority of the bag-of-words baseline, also in absolute terms compared to all participants in INEX 2002, can only be guessed at. The fact that an out-of-the box IR system without any modifications can be in the top 3 of all runs submitted to INEX can be seen as a confirmation of

InQuery's reputation as a high quality IR system. A second factor that may have a positive influence on the results is the particular selection of representations for the main experiments, which were also used in merged form in the bag-of-words baseline. Out of the ten representations, especially the cited titles, abstracts, identifiers and article titles did well on their own, as described in Section 7.2.3.2 above. It is by no means certain that using, e.g., the full text of the documents instead, would have resulted in better performance. The selection of representations was guided by the theory of polyrepresentation as described above, and this may have served to reduce noise compared to the full text as used by the main part of the participants in INEX 2002. This was, however, not tested in the dissertation.

The fine performance of the bag-of-words baseline in INEX 2002 also points to a characteristic of joint IR initiatives: The first year(s) is usually spent tuning the systems to the particular IR tasks without any noteworthy improvements on earlier approaches. Therefore, real improvements in performance are often not obvious until the experiences gained in the first year(s) has been implemented into new IR techniques or the existing ones have been modified.

The remainder of the research questions represent attempts to gain such experiences with the best match boomerang effect and the polyrepresentation baseline by studying the effect of the variables on performance. In this sense the experiments represent what has been characterised as "knob twiddling". When one interprets the results in relation to the INEX 2002 results it must be kept in mind that any of the other INEX participants could have done the same kind of post hoc optimisation with their IR techniques.

### *8.3.2 Research question 2*

Research question 2 investigates if there were significant differences in IR characteristics and performance between the ten individual representations of the scientific full text documents. Differences among the representations are interesting, because such differences may need to be taken into account in future studies that incorporate implementations of the theory of polyrepresentation. In the present dissertation all indexes, except the two citation indexes, were built from the individual representations, and they also formed the basis of the polyrepresentation baseline and the best match boomerang effect. Note that the representations were indexed in *separate* InQuery databases as described in Section 6.2.5 above. Quite different results might

have been obtained if all representations were indexed as fields in a single database, because the *idf* values would have been the same across all representations (See Section 6.2.1 and Formula 2). When separate databases are constructed as in the main experiment, the *idf* values are solely dependent on the particular representation indexed in each database. The latter case is probably better suited for implementations of the theory of polyrepresentation. This was not tested in the dissertation, but is an obvious research question for future research involving polyrepresentation.

As described in Section 7.2.3.2 above, a top group of four representations showed significant differences to the lowest performing representations, but rarely differences among themselves. The top group consisted of two functional (author generated) representations, the article titles and the abstracts. In addition it held two representations that are, at least partially, generated by other cognitive agents, the cited titles and the identifiers. The article titles and abstracts have traditionally been used to represent scientific documents in IR experiments, and therefore it is not surprising that they performed well because of their rhetoric function in the articles. Given that the corpus covers a scientific and technical discipline, the article titles and abstracts can be expected to summarise the contents fairly precisely and are not likely to be metaphorical as it is more often seen in the Humanities.

From a cognitive perspective the high performance of the cited titles and identifiers are not surprising either, as other cognitive agents than the author are involved in their generation. The cited titles are unique in that they may be regarded as being both functional and cognitive, because the author selects which references to include, but the cited titles themselves are generated by a different cognitive agent – the original author. The cited titles displayed the highest absolute AvgP scores of the ten representations (See Table 7.10 and Figure 7.5), and although no significant differences could be identified to the other three in the top group, the cited titles had the largest number of difference to the remaining six representations. This top placement among the ten representations is remarkable, however, when considering the results obtained by Salton and Zhang (1986), who could not find a method to exploit cited titles that would consistently yield good results. There are two main differences between their study and the use made of cited titles in this dissertation:

*First*, Salton and Zhang merged the extracted terms with the other index terms taken from article titles and abstracts, hereby expanding the representation of the documents instead of using them as a separate representation.

*Secondly*, Salton and Zhang used the CACM and ISI test collections, in which cited titles could only be identified for those references that were at the same time part of the test collections as source documents. That is, if a referred document was not also included as a document in the test collection its cited title could not be extracted for the study[70]. They do not present exact statistics on how large a proportion of the references from the bibliographies that could be identified within the test collections, only on the proportion of documents whose representations were altered by the cited titles (25 to 40 % of the documents). In contrast, the cited titles extracted from the INEX corpus include the titles of all cited journal articles, conference papers, and named book chapters (84 % of all references) as described in Section 6.2.4 above. Only references that did not contain the `<atl>` tag were not included. These consist mainly of references to whole books and whole reports.

The fact that the identifiers show significant differences to the worst performing representations, and that the descriptors do not, is interesting, although there was no direct statistical difference between the two representations. Both are cognitive representations generated by the same indexer, but with very different methods: The descriptors are controlled terms that have to be taken from the INSPEC thesaurus according to its rules, whilst the identifiers are uncontrolled words and phrases freely chosen by the indexer. The slightly better performance of the identifiers can be explained by the fact that relatively few descriptors are assigned: there were a total of 57,495 index keys in the descriptor representation versus 166,423 index keys in the identifier representation (Table 6.3). In addition, the descriptors are taken from an artificial language that is less likely to occur in the query keys even after stemming. Automated query expansion techniques such as those investigated by Kekäläinen (1999) could probably be used to increase the performance of the descriptor representation considerably. All the representations might benefit from this kind of expansion, e.g., the

[70] This limitation also holds for ISI's extraction of Keywords Plus. However, chances are much greater that a given reference is also a source item in ISI's citation indexes simply because of their size.

query keys could be expanded by using the INSPEC thesaurus and the expanded queries run against any of the representations. However, the overall effect of this is hard to predict when working with many different representations as in the present dissertation, and much work may have to go into adapting the expansion techniques to each representation as experienced by Madsen and Pedersen (2003).

The table captions and conclusions were the worst performing of the representations, and all four representations in the top group performed significantly better than them. In addition, the cited titles, abstracts and identifiers performed significantly better than the author keywords. Common to all three representations was that they occurred only in 23 to 32 % of the documents in the corpus (See Table 6.3), which may in part explain their low performance. Furthermore, the author keywords and table captions had some of the lowest numbers of index keys, both unique index keys and in total (Table 6.3), which reduces the likelihood of matches with the query keys. Indeed, these two representations accounted jointly for more than half of the cases where fewer documents than the DCV_step1 threshold were retrieved in Step 1 of the best match boomerang effect (Table 7.3).

Conclusions could only be identified in 23 % of the documents in the corpus by the simple parser used as described in Section 6.2.2 above, which to a certain extent may account for the low performance obtained with the representation. However, it was possible to retrieve the maximum number of documents needed, except with the conclusions in a few cases (Table 7.3). Although the introductions displayed higher performance in terms of absolute AvgP scores than the conclusions, they were not significantly different from any of the other representations, except for the cited titles which performed significantly better than the introductions. The figure captions have the same lack of significant differences, although they occur in 67 % of the documents in the corpus and have high numbers of index keys, both unique and in total (Table 7.3). Thereby, except for the cited titles, all the alternative functional representations (introductions, conclusions, figure captions and table captions) displayed either inferior or indistinguishable performance from the rest of the representations. For future experiments these representations may either be left out entirely, or they could alternatively be merged with other representations of similar types, for instance, the author keywords might be merged with the article titles, and the table captions with the figure captions. It is noteworthy that the introductions and the conclusions were not

included in the top group. This contrasts with the results of Lahtinen, who found that the first sentence of the first paragraph, and the last sentence of the last paragraph contained a higher proportion of good index terms (2000, p. 139-143). The reason for this difference may be that whole introductions and conclusions were used in the main experiments. The introductions was by far the largest representation when measured in the total number of index keys (Table 6.3), and the rest of the introductions, apart form the first sentence, may have introduced too much noise compared to, e.g., the article titles and the abstracts. However, another and more likely reason behind their low performance lies in the fact that the introductions and conclusions could only be identified in a limited part of the corpus (and only in the Transactions). A closer analysis of the recall base could determine if this is the case, for instance by examining the proportion of relevant documents between the Transactions and Magazines.

### *8.3.3 Research questions 3 and 4*

Research questions 3 and 4 both investigated the effect on performance of increasing the number of documents from which references were extracted in Step 1 of the best match boomerang effect. Research question 3 investigated this in relation to the polyrepresentation baseline, and research question 4 in relation to the best match boomerang effect. The two questions are treated together here because the same sets of documents were used as basis for both types of runs, and the DCV_step1 threshold controlled the size of these sets.

As discussed in Section 7.2.2, the polyrepresentation baseline cannot in theory retrieve 100 documents (the DCV_step3 threshold used in all runs) for values of DCV_step1 lower than 10, and that in practice this happened at a value of DCV_step1 between 32 and 64 (Table 7.5). Therefore, the differences to runs at values lower than DCV_step1 = 32 described in Section 7.2.3.3 are statistically valid, but much lower performance can be expected with great certainty for these runs. Thus, for research question 3 it is evident that performance will rise from the lowest values of DCV_step1 up to a certain point. The interesting aspect of research question 3 is then whether performance continued to rise as the DCV_step1 was increased, or if performance began to decrease, and if any such differences were statistically significant. The absolute AvgP scores over the range of DCV_step1 values did indeed increase up to a peak at DCV_step1 = 64, and then decreased again (See Table 7.6 to Table 7.9). It is interesting to note that the drop in AvgP from DCV_step1 = 64 to 128 was much greater than the drop to

DCV_step1 = 32 for all four quantification functions. This suggests that the performance of the polyrepresentation baseline followed the DCV_step1 threshold closely, and dropped as soon as 100 documents were retrieved. Thus, instead of providing a more firm statistical basis for the final ranking in the polyrepresentation baseline, a larger number of documents introduced noise and reduced the quality of the final ranking. This can be seen as a consequence of an implementation of the theory of polyrepresentation that was at the unstructured end of the polyrepresentation continuum: Because a document was retrieved if it contained just one of the query keys, the lowest ranking in the output from InQuery would mostly be of little relevance to the information need. The performance of the polyrepresentation baseline therefore decreased as soon as 100 documents had been retrieved. In addition to using more structure in the queries, the polyrepresentation baseline could be given more structure by requiring that a document must occur in a minimum number of representations (e.g., 5), or that it should appear in at least one functional and one cognitive. Any of these approaches would alone decrease the number of documents available for forming the final rank in the polyrepresentation baseline, and in some cases the consequence may be that too few documents can be retrieved (e.g., to fulfil the requirement that a minimum of 100 documents must be retrieved in order to obtain optimal scores with the performance measures as in the main experiment). This problem was experienced by Madsen and Pedersen with a highly structured implementation of the theory of polyrepresentation (2003). They solved the problem by expanding the query terms from a domain thesaurus.

The statistical tests of research question 3 showed that the middle range of DCV_step1 from 16 to 128 had few significant differences among themselves, but many significant differences to the lower and higher values. That is, increasing the DCV_step1 value up to the 16 to 128 interval did improve performance of the polyrepresentation baseline, and any further increases resulted in a decrease of performance. It is noteworthy that DCV_step1 = 16 was included in this interval. As discussed above, it could be expected that this run would display significantly poorer performance than the best runs, as it did not on average retrieve all the 100 documents used for calculating the performance measure. This points to two characteristics of the performance measure and the statistical test used in the main experiment:

*Firstly*, on one hand the failure to identify a significant difference between the runs with DCV_step1 at 16 and, e.g., 32 or 64 may indicate that the Friedman test as used in the analysis of the main experiment was rather conservative. A more powerful version of the Friedman tests is given in Hull (Hull, 1993) which might have been able to distinguish better between the runs. Unfortunately this version was not offered by the used statistical software (SPSS), and time was not available to construct a program that could carry it out.

*Secondly*, on the other hand the AvgP measure may have functioned very well in the sense described in Section 6.4.1, by rewarding the IR techniques that retrieved highly relevant documents early in the rank, and by giving only moderate reward for relevant documents retrieved late in the rank. The polyrepresentation baseline runs carried out with a not too low a DCV_step1 value would then not have been punished drastically, even if they did not retrieve all 100 documents, as long as they retrieved a certain proportion of relevant ones among the ones they did retrieve. Since the results obtained with the polyrepresentation baseline were very closely related to the number of documents used to compute the AvgP scores, more knowledge of its performance could probably be acquired by studying (nD)CG curves of the polyrepresentation baseline.

Research question 4 investigated if performance of the best match boomerang effect could be improved by increasing the number of documents from which references were extracted, that is, by using higher values of the DCV_step1 threshold. As discussed in Section 7.2.1, the DCV_step1 had a great influence on the number of seed documents in Step 3 of the best match boomerang effect: 58 seed documents were used per query on average at DCV_step1 = 2, and 10,486 at DCV_step1 = 512 (Table 7.4). The statistical tests of research question 4 in Section 7.2.3.4 showed that there were few significant differences between the runs in the 4 to 128 range of DCV_step1, and that significant differences were mainly found between these runs and the runs made with the highest values of the DCV_step1 threshold (256 and 512). The generalized functions also found many significant differences between the 8 to 64 DCV_step1 range and the runs made with the lowest value of DCV_step1 = 2. That is, runs made with DCV_step1 values in the 2 to 128 interval performed significantly better when measured with the strict functions. The same was true for the generalized function except for DCV_step1 = 2, which displayed significantly inferior performance compared to most of the runs in the 8 to 64 interval of DCV-step1. The latter result is most likely a consequence of the fact

that significantly less than 100 documents were identified in Step 3 at DCV_step1 = 2 (Table 7.5). Any further increases to DCV_step1 at 256 or 512 resulted in decreasing performance for all functions. The large interval on the DCV_step1 range without any significant differences is remarkable, because there were so great differences in the number of seed documents used in the forward chaining. For instance, more than ten times as many seed documents were used in the forward chaining on average at DCV_step1 = 4 than at DCV_step1 = 64 for all three values of CCV_step2 (Table 7.4).

The results show that there was not much difference in performance between extracting citations for Step 2 from relatively few documents (say 4 or 8) to extracting them from quite a lot of documents (say 64 or 128). This indicates that the best match boomerang effect benefits from restricting the source documents from which references are extracted to those that were ranked highest by InQuery. Because the best match boomerang effect was at the unstructured end of the polyrepresentation continuum, as discussed in Chapter 3, no attempts were made to impose structure on the queries in order to ensure that the documents used as sources for the best match boomerang effect did indeed contain all aspects of the information needs. Instead, InQuery's ranking mechanism was relied on to provide a high quality input, which would hopefully ensure that the overlaps between representations were genuine overlaps and not just random matches across the set. The results of research question 4 show that on average InQuery was successful in providing a sufficiently high quality input for values of DCV_step1 up to 128. The inferior performance displayed above this threshold is probably caused by too low quality overlaps. That only very few significant differences could be identified within this interval may be because the DCV_step1 threshold is rather crude: It reduces the number of source documents in Step 1 to a fixed number irrespective of the characteristics of individual topics and representations. A more sophisticated threshold would probably result in better performance with the best match boomerang effect. For instance, instead of a fixed number of documents, the DCV_step1 threshold could be adapted to the individual topics and representations by reducing the document sets in Step 1 to a certain *percentage* of the top ranked documents returned by InQuery, or to documents that InQuery rank above a certain belief value (See Section 6.2.1). The strict structures created with the queries in Madsen and Pedersen's study represent an attempt to ensure high quality of the overlaps, but at the structured end of the polyrepresentation continuum. Another cause for the fact that only few differences

could be identified may be the possibility of interactions with the other variables. Such interactions were not studied in research question 4, but are given attention below.

### *8.3.4 Research question 5*

Research question 5 investigated if increased performance could be obtained by reducing the number of seed documents in the weighted query in Step 3 of the best match boomerang effect to the seed documents with the highest weights. By this limitation, some of the uncertainty introduced by the automatic translation of the information needs into weighted seed documents might be reduced because only the strongest evidence would be used.

Three values of the CCV_step2 value were tested in the main experiment as described in Section 6.3.2 above. The statistical tests showed that there were very few differences between the three levels: only 11 % of the 324 possible differences were significant. That is, overall it did not make a difference on performance when the number of seed documents was reduced to the ones with the highest weights. A few tendencies can be observed in the small number of differences that were statistically significant (See Table 7.14), which indicate interactions with the other variables. The differences tended to be either at the two extremes of the DCV_step1 range, or in the middle of it. The differences at the highest values of DCV_step1 (256 and 512) all occurred when the flat citation index were used for the extraction of citation to Step 2 (ff or fx). The runs at these DCV_step1 values were found to perform significantly inferior to the other runs in research question 4 above. The differences in the low (2 and 4) and middle (16 to 64) ranges all occur when the expanded citation indexes were used in Step 3 (xf or xx), except for a single case. The runs at these DCV_step1 values were found to perform significantly better in research question 4. What to deduce from these interactions is not immediately obvious, and more advanced statistical tools may be needed to study the interaction between the variables. The CCV_step2 threshold represents an approach that is at the unstructured end of the polyrepresentation continuum. It is not as rigid as the DCV_step1 threshold because the CCV_step2 threshold was implemented as a percentage, and not a fixed number of items. Therefore, no immediate possibilities present themselves for improving the threshold. Instead, more structured approaches, closer to the structured end of the polyrepresentation continuum, could be investigated in future research. For instance, it could be required that the citations must appear in

more than two pools in Step 2, or in certain types of pools, e.g., simultaneously in pools extracted from a functional representation and a cognitive representation.

### *8.3.5 Research questions 6 and 7*

Research questions 6 and 7 both investigated if better performance could be obtained by using expanded citation indexes instead of conventional flat citation indexes in the best match boomerang effect. Research question 6 investigates this in relation to the extraction of citations for Step 2, and research question 7 investigates it in relation to the index against which the weighted sum citation queries were run in Step 3. As proposed by Herlach (1978), the exploitation of expanded citation indexes for IR purposes might improve performance because multiple mentions of the same reference in the full text might point to references that are more central to the theme of the document.

The statistical tests showed a similar pattern for both research questions. Out of the possible differences, 39 to 54 % were significant, and the expanded citation index did not perform better in a single of these cases (Table 7.15 and Table 7.16). There was a clear interaction with the DCV_step1 variable in the results: The vast majority of the runs in the 32 to 512 range of the DCV_step1 threshold showed this tendency between the two types of citation indexes: If the str_inex quantification function is ignored the flat citation indexes performed significantly better in 100 % of the runs tested in research question 6, and in 73 % of the runs tested in research question 7. Only few significant differences could be found at values of DCV_step1 below 32 with any of the quantification functions. That is, in relation to research question 6 the extra statistical information offered by the expanded citation index did not affect the weights of the seed documents in such a way that better performance could be obtained. For values of DCV_step1 at 32 and above it even decreased performance significantly. In relation to research question 7 the extra statistical information offered by the expanded citation index did not affect the weights of the document representations in such a way that better performance could be obtained. For values of DCV_step1 at 32 and above it even decreased performance significantly in most cases.

The interaction with the DCV_step1 variable is interesting because it spans the divide that was identified in research question 4 above, where a significant decrease in performance occurred at values of DCV_step1 = 256 and upwards. When the results of

research question 6 and 7 are compared to the number of documents retrieved in Step 3 of the best match boomerang effect over the range of DCV_step1 values (Table 7.5) a different pattern emerges: The runs at DCV_step1 = 32 and above retrieved 100 documents or more in all cases but one, regardless of the citation index used. Therefore, there is a risk that the lack of differences in the 2 to 16 range of the DCV_step1 threshold is an effect of the AvgP measure, because the computation of it was based on 100 documents. More detailed knowledge could possibly be acquired about the effect at lower values of DCV_step1 if other measures like those based on (nD)CG computations were used.

## 8.4 Concluding discussion

On the whole, the factors tested in the experiments with the best match boomerang effect did not influence performance to any great extent, and the absolute performance of both the best match boomerang effect and the polyrepresentation baseline did not at any time exceed that of the bag-of-words baseline. The overall outcome may therefore be regarded as rather discouraging for the use of polyrepresentation in general and the best match boomerang effect in particular. However, a good deal can be learned from the results.

While the best match boomerang effect did by no means perform as well as the conventional bag-of-words baseline, the experiments did show that the best match boomerang effect was able to retrieve relevant documents. For instance, the n(D)CG curves in Figure 7.3.c-d testify that, at its best, the best match boomerang effect (and the polyrepresentation baseline) was able to retrieve 30 to 60 percent of all of the relevant documents in the recall base up to rank 100 compared to 45 to 80 percent for the bag-of-words baseline. It is important to note, that this level of performance was achieved based on a query in natural language without the specification of seed documents in advance, and without any structure imposed on the queries. The only kind of structure was the partition of the document representations into 10 databases and the thresholds used in Step 1 and Step 2 – otherwise the rest of the experimental setup consisted of the usual elements in best match IR experiments. Thereby the experiments with the best match boomerang effect have shown that references and citations *can* form an integrated part of automatic indexing and retrieval techniques on the same terms as the conventional approaches, i.e., without the need for the queries to specified in any

special manner. This is noteworthy because it was not certain that the promising methods from the earlier research reviewed in Chapter 4 would function in a best match environment without specifying good seed documents in advance. Thereby one of the most important obstacles for exploiting references and citations on equal terms with term-based representations in best match IR systems has been removed, because the main experiment with the best match boomerang effect demonstrates that it is possible, at least to some extent, to identify good seed documents automatically. That the performance of this first attempt does not reach or surpass the level of a very well developed and thoroughly tested approach like that implemented in InQuery is not a hindrance for carrying out research that may improve the results. A number of consequences of the experiments in the dissertation are drawn below, and a number of suggestions for such future research with the best match boomerang effect are considered based on the experiences gained from the experiments.

The best match boomerang effect as tested in the dissertation is fairly complex with many variables and a complicated set of procedures that need to be completed in order to execute it. Since the general idea of extracting references from the top ranked documents returned by a best match system has been shown to work fairly well in a complex setting, future research might investigate how this idea would perform in a simpler setting. As a consequence one fight, for instance, consider testing a modified version of the best match boomerang effect that does not exploit polyrepresentation in Step 1 and Step 2: references could simply be extracted from the top(n) documents returned by the bag-of-words baseline in Step 1 into a single pool, sorted by their frequency of occurrence in the pool, and the top(n) of these could be used as seed documents in a weighted query in Step 3 of the best match boomerang effect. The same thresholds (DCV_step1 and CCV_step2) may be used in such experiments, and the results would be directly comparable to the ones reported in the dissertation. In retrospect, this would have been interesting as a baseline for the experiments in the dissertation, because it could indicate how dependent the best match boomerang effect is on polyrepresentation. On the other hand, it might be considered to work on refining the complex version of the best match boomerang effect. For instance, it could be required that the citations must appear in more than two pools in Step 2 of the best match boomerang effect, or that the citations must appear in certain combinations of pools. Such restrictions would move the best match boomerang effect towards the structured end of the polyrepresentation continuum.

The best match boomerang effect cannot be recommended as an alternative to conventional approaches based on the results of the experiments in the dissertation. Instead the best match boomerang effect might form part of an algorithm in combination with a conventional approach such as the bag-of-words baseline. In a sense this in the spirit of polyrepresentation because the exploitation of references and citations would be one of many different approaches to be combined. An interesting question in relation to this is whether the best match boomerang effect identifies a significant number of documents not found by the bag-of-words baseline and the polyrepresentation baseline? Given the alternative nature of its representation of documents and the translation of queries into seed documents this not an unreasonable expectation. In the pre-experiment reported in Chapter 4 this was indeed the case: a large number of documents not found in Step 1 were identified by the exact match boomerang effect. This was not analysed in the main experiment with the best match boomerang effect because of lack of time, and it is not known if the same is the case in the main experiment. It is recommended for future research that the differences between the documents retrieved by the best match boomerang effect, the bag-of-words baseline, the polyrepresentation baseline, and the 10 individual representations at different levels of DCV_step1 are analysed. Such an analysis could provide indications of whether the best match boomerang effect identifies unique, relevant documents not found or ranked low by other approaches. If this was shown to be the case, the best match boomerang effect might be combined with other approaches, be offered as a way to expand a retrieved set or as a method to retrieve related documents in operational systems. Such implementations might be offered as part of their search engines by the large scientific publishers like Elsevier[71] who have 100,000s of scientific journal articles in XML, or in digital libraries like those discussed in Section 4.4.3.

Scientific articles in XML provide many possibilities for extracting specific information from the XML structure of the articles. A few representations were extracted that are not normally available in documents not formatted in XML: table and figure captions, introductions and conclusions, and the cited titles from the bibliographies. Some knowledge was gained about their IR characteristics and performance individually.

[71] http://www.elsevier.com/

Most notably the cited titles extracted from the bibliographies performed very well – also compared to representations like the articles titles and the abstracts. The rest of the alternative representations did not perform very well on their own. Such knowledge can be very valuable when designing more complex IR approaches, and many more types of representations should be examined. This might lead to the exclusion of certain representations (like the table captions), or merging of others (one might, for instance, consider merging the author keywords with the article titles). This kind of knowledge is especially valuable for experiments with polyrepresentation because high quality input from each representation is needed in order for polyrepresentation to work properly. It may for instance be necessary to modify and adapt the queries to each representation such as done by Madsen and Pedersen (2003). Alternatively the database indexes may be treated differently for each representation, e.g., by using specific stop word lists adapted to each representation. In the pre-experiment the high quality input was ensured by the structure of the Boolean search strings carefully formulated by the test person. In contrast, the experiments with the best match boomerang effect were at the unstructured end of the polyrepresentation continuum. More structure may be added on the query side in future research using, e.g., the principles for structured queries in probabilistic retrieval investigated by Kekäläinen (1999). To aim of using a more structured approach in the treatment of queries would be both to better target them towards individual representations, as well as to give emphasis to overlaps between particular representations.

# 9 Summary and conclusions

The preceding chapters of the dissertation have each contributed to the investigation of references and citations as an integrated part of automatic indexing and retrieval techniques. The theoretical background of the dissertation is the cognitive viewpoint in Information Science and the theory of polyrepresentation as discussed in Chapters 2 and 3. The purpose of this chapter is to summarise the main objectives of the dissertation and to present its major results as well as to recommend directions that may be taken in future research.

## 9.1 Summary of dissertation objectives and results

The main objective of the dissertation is to investigate the use of references and citations as an integrated part of automatic indexing and retrieval techniques operating on scientific full text documents. The main outcome of the dissertation is the boomerang effect, which serves as frame for the investigation of references and citations in IR, and the results obtained with the boomerang effect.

There are two main motivations behind the dissertation: *First*, the few scattered studies from both the system-driven and the user-oriented tradition, that have investigated references and citations for IR purposes, have generally shown promising results. *Secondly*, scientific documents are for the first time becoming available in large quantities in electronic form. This offers new possibilities for combining conventional automatic indexing and retrieval techniques with the exploitation of references and citations in IR.

As stated in Chapter 1 the overall research question which is investigated analytically and empirically in the dissertation is:

> *Which factors affect the behaviour and performance of automatic indexing and retrieval techniques given that references and citations are an integrated part of the document representation of scientific full text documents in the IR system?*

The creator of the citation indexes for science, Eugene Garfield, presented the citation index as "an association-of-ideas index" (1955). The underlying assumption behind this is that there is some kind of semantic relationship between the referring and the cited document. The question is, however, what characterises this semantic link and how it might be exploited for IR purposes. The analytical investigation in the dissertation consists of an analysis in Chapter 4 of references and citations as alternative representations in best match IR, that seeks to investigate *why* and *how* references and citations might be useful in IR, and what factors that might affect the behaviour and performance of best match IR systems with references and citations as alternative representations.

The 'why' is addressed in an analysis of the literature on citer motivations and citation behaviour. According to Cronin (1984) two theoretical positions on the subject may be identified, a *normative* position and a *social constructivist* position. Empirical research tends to give partial support to both positions, and no overall theory has emerged that can explain the observed citation behaviour entirely. References and citations would clearly be useful in IR *if* citation behaviour adheres mainly to the reasons put forward by normative position, but maybe not so useful if it is dominated by the rhetorical-only reasons advocated by the social constructivist position. The most unifying theory considered in Chapter 4 is Small's concept of references as 'concept symbols' which stand for an idea that is being used in the course of an argument (1978). This allows for social or political functions of references, but only within certain bounds because the references have to function as part of a rhetorical context. From the point of exploiting references in IR the idea is attractive: regardless of whether or not some references are omitted, forgotten, biased etc. the references *actually given* function as symbols for a concept. This can explain why references and citations are useful in IR: Because the references represent concepts when used for document representation, they are well-suited for IR purposes as long as the user's information need can be expressed in the same concepts, i.e., as seed documents. In addition, Small's notion of references as concept symbols can explain the variation in performance described below, because a concept can be represented by many different references.

The 'how' is addressed by reviews of earlier studies of references and citations in IR, where a range of different uses have been made of them. As indicated by Figure 2.2 the representations of documents must be matched with compatible representations of the

user's information need in any IR system. The dominant use of references and citations are as seed documents that represent the user's information need in a forward chaining, where documents are retrieved that refer to a given seed document. The main results emerging from the reviewed studies are:

- References and citations are fundamentally different from conventional term-based representations, e.g., they retrieve different documents for the same request, and the overlap between the two is typically small.
- References and citations function as very specific indicators of document content, and tend to increase precision at the expense of recall, and typically show large variation in performance across different information needs.
- The seed documents used to represent the user's information need are crucial.

These points indicate that references and citations have potentials as representations in IR if methods can be found that can improve the selection of seed documents to represent the user's information need. The ability of the seed documents to represent the user's information need is identified as the *main factor* that affects the performance of references and citations as representations in IR. A related factor is that because references and citations tend to be very specific indicators of document content performance might be improved by using a larger number of seed documents than in previous studies.

The seed documents used in the studies reviewed in Chapter 4 were all intellectually selected, either by domain experts or end users. However, very little research has been done on what characterises good seed documents, and how to obtain them from human actors. In addition, the lack of seed documents in the standard test collections used in the system-driven tradition is probably one the main causes of why rather few studies from this tradition have attempted to exploit references and citations.

The boomerang effect proposed in Chapter 5 represents an attempt to solve the problem of obtaining an adequate representation of the information need in the form of seed documents, thus attempting to address the main factor identified in the analytical analysis. The boomerang effect does this by an automatic translation of the information need expressed in natural language into seed documents as described in Chapter 5. Thereby it can use the same representations of the information need as used in conventional IR systems. Candidate seed documents are extracted from sets of documents that are identified by the request in natural language. Inspired by the theory

of polyrepresentation (Ingwersen, 1996) the request is matched against several cognitive and functional representations of the documents, and it is attempted to reduce the uncertainty introduced by the automatic translation by emphasising the extracted seed documents that occur in several of the sets (i.e., in the overlaps between them) as described in Chapter 5. The boomerang effect served as the framework for the experiments. These showed how the identified factors affect the behaviour and performance of references and citations in best match IR, by testing a method for the automatic identification of seed documents, and by implementing the number of seed documents used as a variable. In addition, the implementation of the boomerang effect in the main experiment facilitated the creation of two baselines, one consisting of a simple implementation of polyrepresentation without a citation index, and a conventional bag-of-words baseline.

The idea of a polyrepresentation continuum is proposed as an extension to the theory of polyrepresentation. Two versions of the boomerang effect were tested in the dissertation at each end of the polyrepresentation continuum: a Boolean version closest to the structured pole of the continuum, and a best match version at the unstructured pole.

The Boolean version was tested in the pre-experiment with the purpose of gaining insights into the feasibility of implementing the boomerang effect in a larger experiment. Three real work tasks from a medical researcher were investigated using SCI as database, and the pre-experiment thereby used methods close to those used in the user-oriented tradition to IR research presented in Chapter 2. Two research questions were addressed. The *first* question investigated if the boomerang effect could at all retrieve relevant documents from the network of references and citations without specifying seed documents in advance. The results showed that this was possible, and the boomerang effect had an expansion effect compared to the conventional retrieval method. The *second* question investigated if a larger proportion of relevant documents could be identified in the overlaps identified as result of the strategy. This would be expected from the theory of polyrepresentation, and was indeed shown to be the case: there was a clear tendency for a greater proportion of relevant documents at higher overlap levels. Most of the documents used as source documents for extracting references were also found by the boomerang effect and ordered in the overlap structure. The tendency from the second research question was found both for the source documents and for the extra documents added by the boomerang effect, although

the extra documents displayed lower performance on average. The documents from the highest overlap levels could be shown first to a user, thereby increasing the chance that relevant would be viewed first by the user. Only three work tasks were tested in the pre-experiment and no statistical testing could be carried out. The conclusions should therefore be interpreted with caution.

The best match version of the boomerang effect was implemented and tested in the main experiment using the INEX test collection, along with the two baselines as described in Chapter 5. All the conclusions presented here are therefore dependent on the particular characteristics of the INEX test collection, and because no other test collections were available as control for the obtained results it is not known to what extent the results may be generalised to other corpora.

The main experiment was carried out within the system-driven tradition presented in Chapter 2. Prior to the main experiment a preliminary version of the best match boomerang effect was submitted as an official run to INEX as part of the requirement of getting access to the test collection. Relatively inferior performance was experienced with the best match boomerang effect and the polyrepresentation baseline in the official INEX results. This spurred more detailed studies of the variables of the best match boomerang effect in the main experiment as discussed in Chapters 7 and 8. A post-hoc failure analysis of the official INEX runs showed that the best match boomerang effect and the polyrepresentation baseline were both very sensitive to the quality of the source documents used as input in both. The best match boomerang effect in particular was sensitive to the number of source documents from which references were extracted.

These results led to the implementation of a number of variables and the tuning of these in the main experiment, with the purpose of gaining more knowledge about the factors affecting the best match boomerang effect and polyrepresentation baseline. The variables reflect the second factor identified in the analytical analysis, i.e., that a greater number of seed documents than previously applied in IR research involving references and citations might improve performance. Three variables were implemented to study different aspects of this: a DCV_step1 variable that controlled the number of sources documents from which references were extracted for the best match boomerang effect; a CCV_step1 variable that made it possible to limit the seed documents in the best match boomerang effect to the ones with the highest weights; and a variable that allows for the

use of a either a conventional, *flat* citation index or an *expanded* citation index. The latter contains the references from each document as a flat citation index, and lists in addition the number of times each reference is mentioned in the full text. This is one of the new possibilities offered by scientific documents in full text. The expanded citation index may be used as an alternative to the flat both for the extraction of references in Step 2, and as the citation index against which the seed documents were matched in Step 3 of the best match boomerang effect.

Seven research questions were investigated in the main experiment. The results of these are presented in Chapter 6, and discussed in Chapter 7. The main results are presented below.

*Research question 1* investigated whether the best match boomerang effect could achieve a similar level of performance to what was obtainable with the bag-of-words baseline and the polyrepresentation baseline. This was investigated by testing two of the best performing runs of the best match boomerang effect and polyrepresentation baseline against the bag-of-words baseline. The statistical tests showed that there was no significant difference to the polyrepresentation baseline. This was also obvious from the performance curves where the two were intertwined at several points, although the polyrepresentation baseline was placed slightly better. The bag-of-words baseline was significantly better than both. This was also obvious from the (nD)CG curves, where the bag-of-words baseline was clearly above the best match boomerang effect and polyrepresentation baseline at all points. In conclusion, the optimisation of the variables did result in significant improvements over the official INEX submission, but not significant enough to reach the performance of the bag-of-words baseline. This must be seen in relation to the fact that the bag-of-words baseline performed very well in the INEX results as a whole, where it ranked in the top three among all submitted runs. The best performing best match boomerang effect and polyrepresentation baseline runs in the main experiment did fairly well in comparison to the official INEX results. Their performance is not sufficiently inferior to warrant the conclusion that they must be abandoned entirely. The remainder of the research questions investigated different aspects of the best match boomerang effect and the polyrepresentation baseline to gain knowledge of factors that affect their behaviour and performance.

*Research question 2* investigated if there were significant difference in IR characteristics and performance between the ten cognitive and functional representations used in the main experiment as a background for the analysis of the rest of the results. A top group of four representations showed significant difference to the rest: the article title, abstracts, identifiers and cited titles. In terms of absolute performance the cited titles showed the best performance of the ten individual representations in the main experiment. The latter is remarkable because the earlier study by Salton and Zhang (1986) could not identify a method to exploit cited titles that would yield good results consistently. The main difference to earlier studies is that the cited titles of *all* references were available in the main experiment – Salton and Zhang had only access to those cited titles that also were source documents in the corpus. Apart from the cited titles the alternative representations extracted did not show superior performance: the figure captions and introduction sections showed very few differences to any of the other representations, and the table captions and conclusions showed inferior performance to all in the top group. Part of the reason behind this may be explained by the fact that the latter representations could only be identified in a limited number of the documents.

*Research questions 3 and 4* both investigated the effect of increasing the number of source documents from which references were extracted for the best match boomerang effect (the DCV_step1 threshold). Research question 3 did this in relation to the polyrepresentation baseline. The results showed that the lowest values of DCV_step1 resulted in lower performance. This was not unexpected because the descriptive results showed that the polyrepresentation baseline could not retrieve as many documents as was used to calculate the performance measure at low values of DCV_step1. The results also showed significantly inferior performance at the highest values of DCV_step1, indicating that the extra source documents added acted as noise and resulted in decreasing performance of the polyrepresentation baseline. Similar results were obtained for research question 4, where significantly inferior performance only occurred at the highest and lowest values of DCV_step1. This is remarkable because there was a large difference in the number of seed documents depending on the value of DCV_step1.

*Research question 5* investigated the effect of reducing the seed documents to the ones with the highest weights (the CCV_step2 threshold). Overall, the statistical showed that

there were very few significant differences between reducing the seed documents to the highest weighted ones, and not doing so. The few instances where there were significant differences did not show any clear patterns, but rather indicated that there might be interactions with the other variables.

*Research questions 6 and 7* investigated the effect of using the flat or the expanded citation index in the best match boomerang effect. Research question 6 investigated this in relation to the extraction of citations for Step 2, and research question 7 investigated it in relation to the citation index against which the weighted seed documents were run in Step 3 of the best match boomerang effect. The results were very similar for both questions: Significant differences were found in favour of the flat citation index at values of DCV_step1 larger than 32, and few difference were found below this value. That is, in no cases did the expanded citation index result in significantly better performance over the flat citation index. Higher values of DCV_step1 the expanded citation index even resulted in decreasing performance.

In summary, the following has been learned from the experiments: Automatic translation of requests in natural language into seed documents that can be matched against a citation index is possible with the approach taken inspired by the theory of polyrepresentation (Ingwersen, 1996). The resulting implementation, the boomerang effect, performs as well as a baseline based on the same theory without a citation index, but not as well as a bag-of-words baseline representing a conventional best match approach. A number of factors were examined for their affect on the performance of the best match boomerang effect:

*First*, the number of source documents from which citations were extracted did not influence performance noticeably, except if it was very low or very high (research question 4).

*Second*, the number of seed documents used in the forward chaining did not influence performance noticeably when the seed documents with highest weights were used, except if the number of seed documents was very low or very high (research question 5). Taken together with the first point, it may be concluded that seed documents may be extracted from a fairly low number of source documents, and that a fairly low proportion of the highest weighted seed documents are *sufficient* to provide the best

obtainable performance, although significantly more does not harm performance. There were, therefore, diminishing returns in using greater numbers of both quantities. It should be noted, however, that the number of seed documents necessary was far beyond what a user could normally be expected to provide intellectually. Even if such amounts of seed documents could be supplied it would be very hard utilise them in practice if this was not automated.

*Third*, the use of an expanded citation index both for the extraction of citations and as the database against which to match the seed documents did not result in significantly better performance in a single case, and decreased performance when citations were extracted from many source documents. No significant differences were found, however, between a flat and an expanded citation index in the range that were shown to be sufficient as discussed above.

The limited effect of the factors on the performance of citation searching as implemented in the boomerang effect may be due to the fact that the best match boomerang effect is at the un-structured pole of the polyrepresentation continuum. As indicated by the pre-experiment and Madsen and Pedersen's study (Madsen and Pedersen, 2003) a more structured approach in both step1 and Step 3 of the best match boomerang effect may be beneficial.

The main contributions of the dissertation are:

a) That good seed documents are crucial in any citation search strategy based on forward chaining;
b) The knowledge gained of the limited influence of the factors on the performance of citation searching as implemented in the boomerang effect;
c) The proposal of a polyrepresentation continuum with a structured and an un-structured pole emphasising the range of possible polyrepresentative opportunities along the continuum;
d) That the theory of polyrepresentation did not perform as well as conventional methods when implemented at the unstructured end of the polyrepresentation continuum;
e) The indications that a more structured approach might be beneficial in future research; and finally

f) The verification that the conventional bag-of-words approach remains a very strong model for IR.

Recommendations for future research includes the use of structured queries in Step 1 of the best match boomerang effect and the polyrepresentation baseline; more structure in Step 2 of the best match boomerang effect, e.g., by requiring that the citation must appear in more than two pools, or in specific combinations of pools, as well as the use of more refined representations generated from the full text.

# 10 References


[SPPS Inc.] ([2002]): *NPAR tests*. [Chicago]: [SPSS Inc.]. 21 p. (Chapter from the online documentation for SPSS dealing with nonparametric tests) [http://www.spss.com/tech/stat/Algorithms/11.5/npar_tests.pdf , visited 25-9-2003]

Allan, J., Callan, J., Croft, W. B., Ballesteros, L., Byrd, D., Swan, R. and Xu, J. (1997): Inquery does battle with TREC-6. In: Voorhees, E. M. and Harman, D. K. eds. *The Sixth Text REtrieval Conference (TREC-6)*. [Gaithersburg, MD]: National Institute of Standards and Technology, p. 169-206. (NIST Special Publication ; 500-240) [http://trec.nist.gov/pubs/trec6/papers/umass-trec6.ps.gz, visited 22-8-2003]

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

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. (1979): Information search tactics. *Journal of the American Society for Information Science*, 30(4), 205-214.

Bates, M. J. (1986): Subject access in online catalogs : a design model. *Journal of the American Society for Information Science*, 37(6), 357-376.

Bazerman, C. (1988): *Shaping written knowledge : the genre and the activity of the experimental article in science*. Madison, Wisconsin: University of Wisconsin Press. x, 356 p. ISBN: 0-299-11690-5. [https://wac.colostate.edu/books/bazerman_shaping, visited 10-11-2003]

Belkin, N. J. (1977): Internal knowledge and external information. In: De Mey, M., Pinxten, R., Poriau, M. and VanDamme, F. eds. *International Workshop on the Cognitive Viewpoint*. Ghent: University of Ghent, p. 187-194.

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. (1990): The cognitive viewpoint in information science. *Journal of Information Science*, 16, 11-15.

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 & 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. and Robertson, S. E. (1976): Information science and the phenomenon of information. *Journal of the American Society for Information Science*, 27, 197-204.

Belkin, N. J. and Croft, W. B. (1987): Retrieval techniques. In: Williams, M. E. ed. *Annual review of information science and technology (ARIST), volume 22, 1987*. Amsterdam: Elsevier Science, p. 107-145.

Björneborn, L. and Ingwersen, P. (2001): Perspectives of webometrics. *Scientometrics*, 50(1), 65-82.

Blair, D. C. and Maron, M. E. (1985): An evaluation of retrieval effectiveness for a full-text document-retrieval system. *Communications of the ACM*, 28(3), 289-299.

Bollacker, K. D., Lawrence, S. and Giles, C. L. (1998): CiteSeer : an autonomous web agent for automatic retrieval and identification of interesting publications. In: Syraca, K. P. and Wooldridge, M. eds. *Proceedings of the second international conference on autonomous agents, St. Paul, Minneapolis, USA, May 9-13, 1998*. New York: ACM Press, p. 116-123. [http://citeseer.nj.nec.com/bollacker98citeseer.html (preprint), visited 17-8-2003]

Borgman, C. (1989): All users of information retrieval systems are not created equal : an exploration into individual differences. *Information Processing & Management*, 25(3), 237-252.

Borlund, P. (2000a): *Evaluation of interactive information retrieval systems*. Åbo: Åbo Akademi University Press. vi, 276 p. ISBN: 951-765-049-3. (PhD dissertation)

Borlund, P. (2000b): Experimental components for the evaluation of interactive information retrieval systems. *Journal of Documentation*, 56(1), 71-90.

Borlund, P. (2003): The concept of relevance in IR. *Journal of the American Society for Information Science and Technology*, 54(10), 913-925.

Borlund, P. and Ingwersen, P. (1998): Measures of relative relevance and ranked half-life : performance indicators for interactive IR. In: Croft, W. B., Moffat, A., van Rijsbergen, C. J., Wilkinson, R. and Zobel, J. eds. *SIGIR 98 : proceedings of the 21st annual international ACM SIGIR conference on research and development in information retrieval : August 24-28, 1998, Melbourne, Australia*. New York: ACM Press, p. 324-331.

Bradshaw, S. and Hammond, K. (1999): Constructing indices from citations in collections of research papers. In: Woods, L. ed. *ASIS '99 : proceedings of the 62nd ASIS annual meeting, Washington, DC, October 31-November 4, 1999*. Medford, N.J.: Information Today, [http://dent.infolab.nwu.edu/infolab/papers/paper_desc.asp?ID=10068 (preprint)]

Bradshaw, S. and Hammond, K. (2001): Using citation to facilitate precise indexing and automatic index creation in collections of research papers. *Knowledge-Based Systems*, 14, 29-35.

Bradshaw, S. and Hammond, K. (2002): Automatically indexing documents : content vs. reference. In: Gil, Y. and Leake, D. eds. *Proceedings of the 7th international conference on intelligent user interfaces : January 13-16, 2002, San Francisco, California, USA*. New York: ACM Press, p. 180-181. (Poster paper)

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]

Brookes, B. C. (1975a): The fundamental equation of information science. In: *Problems of Information Science, FID 530*. Moscow: VINITI, p. 115-130.

Brookes, B. C. (1975b): The fundamental problem of information science. In: Horsnell, V. ed. *Informatics 2*. London: Aslib, p. 42-49.

Brookes, B. C. (1977): The developing cognitive viewpoint in information science. In: De Mey, M., Pinxten, R., Poriau, M. and VanDamme, F. eds. *International Workshop on the Cognitive Viewpoint*. Ghent: University of Ghent, p. 195-203.

Brookes, B. C. (1980): The foundation of information science : part I : philosophical aspects. *Journal of Information Science : Principles and Practice*, 2, 125-133.

Brooks, H. (1987): Expert systems and intelligent information retrieval. *Information Processing & Management*, 23(4), 367-382.

Brooks, T. A. (1985): Private acts and public objects : an investigation into citer motivations. *Journal of the American Society for Information Science*, 36(4), 223-229.

Brooks, T. A. (1986): Evidence of complex citer motivations. *Journal of the American Society for Information Science*, 37(1), 34-36.

Bruce, H. (1994): A cognitive view of the situational dynamism of user-centered relevance estimation. *Journal of the American Society for Information Science*, 45(3), 142-148.

Buckley, C. and Voorhees, E. M. (2000): Evaluating evaluation measure stability. In: *SIGIR 2000, Greece*. p. 33-40.

Callan, J. P., Croft, W. B. and Harding, S. M. (1992): The inquery retrieval system. In: Tjoa, A. M. and Ramos, I. eds. *Database and expert systems applications : proceedings of the international conference in Valencia, Spain, 1992*. Wien: Springer-Verlag, p. 78-83. [http://ciir.cs.umass.edu/pubfiles/ir-21.pdf, visited 22-8-2003]

Cawkell, A. E. (1968): Search strategies using the science citation index. In: Houghton, B. ed. *Computer based information retrieval systems*. London: Clive Bingley, p. 49-62. (Reprinted with introduction by Eugene Garfield in: *Essays of an information scientist: 1962-1973, vol. 1*, 1977, p. 49-62.) [http://www.garfield.library.upenn.edu/essays/V1p049y1962-73.pdf (reprint), visited 22-5-2003]

Cawkell, A. E. (1974): Search strategy, construction and use of citation networks, with a socio-scientific example : "amorphus semi-conductors and S. R. Ovshinsky". *Journal of the American Society for Information Science*, March-April, 123-130.

Cawkell, A. E. (1998): Checking research progress on 'image retrieval by shape-matching' using the Web of Science. *Aslib Proceedings*, 50(2), 27-31.

Cawkell, A. E. (2000): Methods of information retrieval using Web of Science: pulmonary hypertension as a subject example. *Journal of Information Science*, 26(1), 66-70.

Chapman, J. and Subramanyam, K. (1981): Cocitation search strategy. In: Williams, M. E. and Hogan, T. H. eds. *National online meeting : proceedings 1981*. Medford, N. J.: Learned Information, p. 97-102.

Chiaramella, Y. (2001): Information retrieval and structured documents. In: Agosti, M., Crestani, F. and Pasi, G. eds. *Lectures on information retrieval : third European summer-school, ESSIR 2000, Varenna, Italy, September 11-15, 2000: Revised Lectures*. Berlin: Springer, p. 286-309. (Lecture Notes in Computer Science ; LNCS 1980)

Christensen, F. H. and Ingwersen, P. (1996): Online citation analysis - a methodological approach. *Scientometrics*, 37(1), 39-62.

Chubin, D. E. and Moitra, S. D. (1975): Content analysis of references : adjunct or alternative to citation. *Social Studies of Science*, 5(4), 423-441.

Cipra, B. (2003): Joint mathematics meetings : diagram masters cry 'Venn-i, Vidi, Vici'. *Science*, 299, p. 651. (in News Focus)

Cleverdon, C. W. (1960): *Report on the first stage of an investigation into the comparative efficiency of indexing systems*. Cranfield: The College of Aeronautics. 166 p.

Cleverdon, C. W. (1962): *Report on the testing and analysis of an investigation into the comparative efficiency of indexing systems*. Cranfield: [The College of Aeronautics]. 312 p.

Cleverdon, C., Mills, J. and Keen, M. (1966): *Aslib Cranfield research project : factors determining the performance of indexing systems : volume 1 : design : an investigation supported by a grant to Aslib by the National Science Foundation*. Cranfield: [College of Aeronautics]. [iv], 120 p.

Cleverdon, C. and Keen, M. (1966): *Aslib Cranfield research project : factors determining the performance of indexing systems : volume 2 : an investigation supported by a grant to Aslib by the National Science Foundation*. Bedford: Published by Cyril Cleverdon. [i-v], 299 p. (test results)

Cole, S. and Cole, J. R. (1987): Testing the Ortega hypothesis: milestone or millstone. *Scientometrics*, 12(5-6), 345-353.

Conklin, J. (1987): Hypertext : an introduction and survey. *IEEE Computer*, 20(9), 17-41.

Cooper, W. S. (1968): Expected search length : a single measure of retrieval effectiveness based on the weak ordering action of retrieval systems. *American Documentation*, 19(1), 30-41.

Cosijn, E. and Ingwersen, P. (2000): Dimensions of relevance. *Information Processing & Management*, 36(4), 533-550.

Cozzens, S. E. (1989): What do citations count? The rhetoric-first model. *Scientometrics*, 15(5-6), 437-447.

Croft, W. B. and Thompson, R. H. (1987): $I^3R$ : a new approach to the design of document retrieval systems. *Journal of the American Society for Information Science*, 38(6), 389-404.

Cronin, B. (1984): *The citation process : the role and significance of citations in scientific communication*. London: Taylor Graham. 103 p. ISBN: 0-947568-01-8.

Damerau, F. (1964): A technique for computer detection and correction of spelling errors. *Communications of the ACM*, 7, 171-176.

De Mey, M. (1977): The cognitive viewpoint : its development and scope. In: De Mey, M., Pinxten, R., Poriau, M. and VanDamme, F. eds. *International Workshop on the Cognitive Viewpoint*. Ghent: University of Ghent, p. xvi-xxxii.

De Mey, M. (1980): The relevance of the cognitive paradigm for information science. In: Harbo, O. and Kajberg, L. 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: Williams, M. E. ed. *Annual review of information science and technology (ARIST), volume 21, 1986*. New York: Knowledge Industry Publications, p. 3-33.

Ding, Y., Chowdhury, G. G., Foo, S. and Qian, W. (2000): Bibliometric information retrieval system (BIRS) : a web search interface utilizing bibliometric research results. *Journal of the American Society for Information Science*, 51(13), 1190-1204.

Duncan, E. B. e. al. (1981): Qualified citation indexing : its relevance to educational technology. In: Duncan, E. B. and McAleese, R. eds. *Information retrieval in educational technology : proceedings of the first symposium on information retrieval in educational technology, 1st April 1981, Aberdeen*. Aberdeen: University of Aberdeen, p. 70-79.

Egghe, L. and Rousseau, R. (1990): *Introduction to informetrics : quantitative methods in library, documentation and information science*. Amsterdam: Elsevier. 450 p. ISBN: 0-444-88493-9.

Egghe, L. and Rousseau, R. (2002): Co-citation, bibliographic coupling and a characterization of lattice citation networks. *Scientometrics*, 55(3), 349-361.

Ellis, D. (1989): A behavioural approach to information retrieval system design. *Journal of Documentation*, 45(3), 171-212.

Ellis, D. (1992): The physical and cognitive paradigms in information retrieval research. *Journal of Documentation*, 48(1), 45-64.

Ellis, D. (1996): *Progress and problems in information retrieval [2nd ed.]*. London: Library Association Publishing. 220 p. ISBN: 1-85604-123-9.

Ellis, D. (1998): Paradigms and research traditions in information retrieval research. *Information Services & Use*, 18, 225-241.

Ellis, D., Wilson, T. D., Ford, N., Foster, A., Lam, H. M., Burton, R. and Spink, A. (2002): Information seeking and mediated searching : part 5 : user-intermediary interaction. *Journal of the American Society for Information Science*, 53(11), 883-893.

Fisher, M. and Everson, R. (2003): When are links useful? experiments in text classification. In: Sebastiani, F. ed. *Advances in information retrieval : 25th European conference on IR research, ECIR 2003, Pisa, Italy, April 14-16, 2003 : proceedings*. Berlin: Springer, p. 41-56. (Lecture Notes in Computer Science ; 2633)

Fox, E. A., Nunn, G. L. and Lee, W. C. (1988): Coefficients for combining concept classes in a collection. In: Chiaramella, Y. ed. *Proceedings of the 11th annual international ACM SIGIR conference on research and development in information retrieval*. New York: ACM Press, p. 291-307.

Fuhr, N., Gövert, N., Kazai, G. and Lalmas, M. (2002): INEX : initiative for the evaluation of XML retrieval. In: Baeza-Yates, R., Fuhr, N. and Maarek, Y. S. eds. *ACM SIGIR'2002 workshop on XML and information retrieval, August 15, Tampere, Finland*. [Tampere]: [SIGIR'2002], p. 62-70. (Workshop notes)

Fuhr, N., Gövert, N., Kazai, G. and Lalmas, M. eds. (2003): *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]. iv, 281 p. (DELOS Network of Excellence on Digital Libraries Workshop Series ; ERCIM -03-W03) [http://www.ercim.org/publication/ws-proceedings/INEX2002.pdf, visited 26-1-2004]

Garfield, E. and Sher, I. H. (1963): *Science Citation Index*. Philadelphia: Institute for Scientific Information. 2672 p. (5 volumes)

Garfield, E. (1955): Citation indexes for science : a new dimension in documentation through association of ideas. *Science*, 122(3139), 108-111. (Reprinted in: *Essays of an information scientist: 1983, vol. 6*, 1984, p. 468-471) [http://www.garfield.library.upenn.edu/essays/v6p468y1983.pdf, visited 26-5-2003]

Garfield, E. (1964): "Science citation index" : a new dimension in indexing. *Science*, 144(3619), 649-654. (Reprinted in: *The awards of science and other essays : essays of an information scientist: 1984, vol. 7*, 1985, p. 525-535) [http://www.garfield.library.upenn.edu/essays/v7p525y1984.pdf, visited 29-4-2003]

Garfield, E. (1965): Can citation indexing be automated? In: Stevens, M. E., Giuliano, V. E. and Heilprin, L. B. eds. *Statistical association methods for mechanized documentation, Symposium proceedings, Washington, 1964.* Washington: National Bureau of Standards, p. 189-192. (National Bureau of Standards Miscellaneous Publication ; 269) (Reprinted in: *Essays of an information scientist: 1962-1973, vol. 1*, 1977, p. 84-90) [http://www.garfield.library.upenn.edu/essays/V1p084y1962-73.pdf ]

Garfield, E. (1970): Trained scientists uses Science Citation Index® to complete bibliographic citations & update searches. *Current Contents*, August 19, p. 119. (Reprinted in: *Essays of an information scientist: 1962-1973, vol. 1*, 1977, p. 119) [http://www.garfield.library.upenn.edu/essays/V1p119y1962-73.pdf, visited 15-6-2003]

Garfield, E. (1979): *Citation indexing : its theory and application in science, technology, and humanities : foreword by Robert K. Merton*. New York: John Wiley & Sons. xi, 274 p. (Information Science Series) ISBN: 0-471-02559-3. [http://www.garfield.library.upenn.edu/ci/title.pdf, visited 5-6-2003]

Garfield, E. (1988): Announcing the SCI compact disc edition : CD-ROM gigabyte storage technology, novel software, and bibliographic coupling make desktop research and discovery reality. *Current Contents*, (22), 3-13. [http://www.garfield.library.upenn.edu/essays/v11p160y1988.pdf, visited 15-6-2003]

Garfield, E. (1990): Keywords plus: ISI's breakthrough retrieval method : part 1 : expanding your searching power on current contents on diskette. *Current Contents*, (32), 3-7. (Reprinted in: *Journalology, keywords plus, and other essays : essays of an information scientist: 1990, vol. 13*, 1990, p. 295-299) [http://www.garfield.library.upenn.edu/essays/v13p295y1990.pdf, visited 18-7-2003]

Garfield, E. (1993): KeyWords Plus(TM) : algorithmic derivative indexing. *Journal of the American Society for Information Science*, 44(5), 298-299. [http://www.garfield.library.upenn.edu/papers/jasis44(5)p298y1993.html (preprint), visited 18-7-2003]

Garfield, E. (1998): From citation indexes to informetrics : is the tail now wagging the dog? *Libri*, 48, 67-80. [http://www.garfield.library.upenn.edu/papers/libriv48(2)p67-80y1998.pdf, visited 26-5-2003]

Geman, S. and Geman, D. (1984): Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6(6), 721-741.

Gilbert, G. N. (1977): Referencing as persuasion. *Social Studies of Science*, 7(1), 113-122.

Giles, C. L., Bollacker, K. D. and Lawrence, S. (1998): CiteSeer : an automatic citation indexing system. In: Akscyn, R. and Shipman, F. M. eds. *Digital libraries '98 : the third ACM conference on digital libraries : June 23-26, 1998, Pittsburgh, PA*. New York: Association for Computing Machinery, p. 89-98. [http://citeseer.nj.nec.com/giles98citeseer.html (preprint), visited 23-4-2003]

Goodrum, A. A., McCain, K. W., Lawrence, S. and Giles, C. L. (2001): Scholarly publishing in the internet age : a citation analysis of computer science literature. *Information Processing & Management*, 37(5), 661-675.

Griffith, B. C., Small, H. G., Stonehill, J. A. and Dey, S. (1974): The structure of scientific literatures II : toward a macro- and microstructure for science. *Science Studies*, 4, 339-364.

Gövert, N. and Kazai, G. (2003): Overview of the initiative for the evaluation of XML retrieval (INEX) 2002. In: Fuhr, N., Gövert, N., Kazai, G. and Lalmas, M. 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: Korfhage, R., Rasmussen, E. and Willett, P. 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.

Harter, S. P. (1986): *Online information retrieval : concepts, principles and techniques*. Orlando, Florida: Academic Press. xi, 259 p.

Herlach, G. (1978): Can retrieval of information from citation indexes be simplified. *Journal of the American Society for Information Science*, 29(6), 308-310.

Hjortgaard Christensen, F. and Ingwersen, P. (1996): Online citation analysis : a methodological approach. *Scientometrics*, 37(1), 39-62.

Hjørland, B. (1993): *Emnerepræsentation og informationssøgning : bidrag til en teori på kundskabsteoretisk grundlag*. Göteborg: Valfrid. 258 p. (Skrifter fron Valfrid ; 4)

Hull, D. (1993): Using statistical testing in the evaluation of information retrieval experiments. In: Korfhage, R., Rasmussen, E. and Willett, P. eds. *SIGIR'93 : proceedings of the 16th annual international ACM SIGIR conference on research and development in information retrieval*. New York: ACM Press, p. 329-338. [http://citeseer.nj.nec.com/hull93using.html (preprint)]

Hutchins, W. J. (1978): The concept of 'aboutness' in subject indexing. *Aslib Proceedings*, 30(5), 172-181.

Ingwersen, P. (1982): Search procedures in the library analysed from the cognitive point of view. *Journal of Documentation*, 38(3), 165-191.

Ingwersen, P. (1984): A cognitive view of three selected online search facilities. *Online Review*, 8(5), 465-492.

Ingwersen, P. (1992): *Information retrieval interaction*. London: Taylor Graham. x, 246 p. ISBN: 0-947568-54-9. [http://www.db.dk/pi/iri/]

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

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: Williams, M. E. ed. *Annual review of information science and technology (ARIST), volume 34, 1999-2000*. Medford, NJ: Information Today, vol. 34, p. 3-51.

Ingwersen, P. (2002): Cognitive perspectives of document representation. In: Bruce, H., Fidel, R., Ingwersen, P. and Vakkari, P. 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 Christensen, F. H. (1997): Data set isolation for bibliometric online analyses of research publications : fundamental methodological issues. *Journal of the American Society for Information Science*, 48(3), 205-217.

Ingwersen, P. and Järvelin, K. ([2004]): *The turn : integration of information seeking and retrieval in context*. Klüwer. (In preparation for 2004)

Ingwersen, P., Larsen, B. and Wormell, I. (2000): Applying diachronic citation analysis to ongoing research program evaluations. In: Cronin, B. and Atkins, H. B. eds. *The web of knowledge : a festschrift in honor of Eugene Garfield*. Medford, New Jersey: Information Today & American Society for Information Science, p. 373-387. (ASIS Monograph Series)

Ingwersen, P., Noyons, E. and Larsen, B. (2001): Mapping national research profiles in social science disciplines. *Journal of Documentation*, 57(6), 715-740.

Institute for Scientific Information (2002): *[Facts about] ISI's Web of Science*. 4 p. (Brochure: Web of Science version 5.0, Rev 4/02) [http://www.isinet.com/presentrep/facts/wosfact5.pdf, visited 4-6-2003]

Järvelin, K., Ingwersen, P. and Niemi, T. (2000): A user-oriented interface for generalised informetric analysis based on applying advanced data modelling techniques. *Journal of Documentation*, 56, 250-278.

Järvelin, K. and Kekäläinen, J. (2000): IR evaluation methods for retrieving highly relevant documents. In: Belkin, N. J., Ingwersen, P. and Leong, M.-K. eds. *SIGIR 2000 : proceedings of the 23rd annual international ACM SIGIR conference on research and development in information retrieval : July 24-28, 2000, Athens, Greece*. New York: ACM Press, p. 41-48. [http://www.info.uta.fi/tutkimus/fire/archive/KJJKSIGIR00.pdf (preprint), visited 3-9-2003]

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

Kaplan, N. (1965): The norms of citation behaviour : prolegomena to the footnote. *American Documentation*, 16(3), 179-184.

Kazai, G., Lalmas, M. and Reid, J. (2003): Construction of a test collection for the focussed retrieval of structured documents. In: Sebastiani, F. ed. *Advances in information retrieval : 25th European conference on ir research : ECIR 2003 : Pisa, Italy, April 14-16, 2003 : Proceedings*. Berlin: Springer-Verlag, p. 88-103. (Lecture Notes in Computer Science ; LNCS 2633)

Keen, E. M. (1992): Presenting results of experimental retrieval comparisons. *Information Processing & Management*, 28(4), 491-502.

Kekäläinen, J. (1999): *The effects of query complexity, expansion and structure on retrieval performance in probabilistic text retrieval*. Tampere: Department of Information Studies, University of Tampere. 141 p. (Acta Universitatis Tamperensis ; 678) ISBN: 951-44-4596-1. (PhD dissertation) [http://www.info.uta.fi/tutkimus/fire/archive/QCES.pdf, visited 10-5-2003]

Kekäläinen, J. and Järvelin, K. (2002a): Evaluating information retrieval systems under the challenges of interaction and multidimensional dynamic relevance. In: Bruce, H., Fidel, R., Ingwersen, P. and Vakkari, P. 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. 253-270.

Kekäläinen, J. and Järvelin, K. (2002b): Using graded relevance assessments in IR evaluation. *Journal of the American Society for Information Science and Technology*, 53(13), 1120-1129.

Kessler, M. M. (1963): Bibliographic coupling between scientific papers. *American Documentation*, (January), 10-25.

Kessler, M. M. (1965): Comparison of the result of bibliographic coupling and analytic subject indexing. *American Documentation*, 16(3), 223-233.

Kochen, M. (1974): *Principles of information retrieval*. Los Angeles: Melville Publishing Company. 203 p. (Information Science Series) ISBN: 0-471-49697-9.

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.

Kwok, K. L. (1975): The use of titles and cited titles as document representations for automatic classification. *Information Processing & Management*, 11(8-12), 201-206.

Kwok, K. L. (1984): A document-document similarity measure based on cited titles and probability theory, and its application to relevance feedback retrieval. In: van Rijsbergen, C. J. ed. *Research and development in information retrieval : proceedings of the third joint BCS and ACM symposium, King's College, Cambridge, 2-6 July 1984*. Cambridge: Cambridge University Press, p. 221-231.

Kwok, K. L. (1985a): A probabilistic theory of indexing and similarity measure based on cited and citing documents. *Journal of the American Society for Information Science*, 36(5), 342-351.

Kwok, K. L. (1985b): Experiments with cited titles for automatic document indexing and similarity measure in a probabilistic context. In: Tauge, J. M. ed. *Proceedings of the 8th annual international ACM SIGIR conference on Research and development in information retrieval : Montreal, Canada*. New York: ACM Press, p. 165-178.

Lahtinen, T. (2000): *Automatic indexing : an approach using an index term corpus and combining linguistic and statistical methods*. Helsinki: Department of General Linguistics, University of Helsinki. [iv], 224 p. ISBN: 951-45-9640-4. (PhD dissertation) [http://ethesis.helsinki.fi/julkaisut/hum/yleis/vk/lahtinen/, visited 22-8-2003]

Lancaster, W. F. (1969): Medlars : report on the evaluation of its operating efficiency. *American Documentation*, 20, 119-142.

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. and Ingwersen, P. (2001): Synchronous and diachronous citation analysis for information retrieval : generating a boomerang effect form the network of scientific papers. In: Davis, M. and Wilson, C. S. eds. *Proceedings of the 8th international conference on scientometrics & informetrics, ISSI2001, Sydney, July 16-20, 2001*. Sydney: The University of New South Wales, p. 355-368.

Larsen, B. and Ingwersen, P. (2002): The boomerang effect : retrieving scientific documents via the network of references and citations. In: Beaulieu, M., Baeza-Yates, R., Myaeng, S. H. and Järvelin, K. eds. *Proceedings of SIGIR 2002 : the twenty-fifth annual international ACM SIGIR conference on research and development in information retrieval, August 11-15, 2002, Tampere, Finland*. New York: ACM Press, p. 397-398. (Poster paper) [http://www.db.dk/blar (Preprint), visited 3-8-2003]

Latour, B. and Woolgar, S. (1979): *Laboratory life : the social construction of scientific facts*. Beverly Hills: Sage Publications. 271 p. (Sage library of social research ; 80) ISBN: 0-80390993-4.

Lawrence, S., Giles, C. L. and Bollacker, K. D. (1999): Autonomous Citation Matching. In: Etzioni, O., Müller, J. P. and Bradshaw, J. M. eds. *AGENTS '99. Proceedings of the Third Annual Conference on Autonomous Agents, May 1-5, 1999, Seattle, WA, USA*. New York: ACM Press, p. 392-393. (Poster paper) [http://citeseer.nj.nec.com/lawrence99autonomous.html (preprint), visited 16-8-2003]

Levenshtein, V. I. (1965): Binary codes capable of correcting deletions, insertions and reversals. *Doklady Akademii Nauk SSSR*, 163(4), 845-848. (Also published in: Soviet Physics Doklady, 1966, 10(8), 707-710)

Lin, X., White, H. D. and Buzydlowski, J. (2003): Real-time author co-citation mapping for online searching. *Information Processing & Management*, 39(5), 689-706.

Lipetz, B.-A. (1965): Improvement of the selectivity of citation indexes to science literature through inclusion of citation relationship indicators. *American Documentation*, 16(2), 81-90.

Luhn, H. P. (1960): Keywords in context index for technical Literature. *American Documentation*, XI(4), 288-295.

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

MacRoberts, M. H. and MacRoberts, B. R. (1986): Quantitative measures of communication in science : a study of the formal level. *Social Studies of Science*, 16(1), 151-172.

MacRoberts, M. H. and MacRoberts, B. R. (1989): Problems of citation analysis : a critical review. *Journal of the American Society for Information Science*, 40(5), 342-349.

Madsen, M. and Pedersen, H. (2003): *Polyrepræsentation som IR metode : afprøvning af polyrepræsentationsteorien i et best match IR system [Polyrepresentation as IR method : test of the theory of polyrepresentation in a best match IR system]*. [Copenhagen]: Danmarks Biblioteksskole. 106 p.+ XLIII p. (In Danish - unpublished MLIS thesis)

Mark Pejtersen, A. (1989): A library system for information retrieval based on a cognitive task analysis and supported by an icon-based interface. *ACM Sigir Forum*, June, 40-47.

Marshakova, I. V. (1973): A system of document connection based on references. *Nauchno-Tekhnicheskaya Informatsiya Seriya 2 : Informatsionnye Protsessy I Sistemy [Scientific and Technical Information Serial of VINITI 6]*, 6(2), 3-8.

Martino, J. P. (1971): Citation indexing for research and development management. *IEEE Transactions on Engineering Management*, EM-18(4), 146-151.

McCain, K. W. (1989): Descriptor and citation retrieval in the medical behavioral sciences literature: Retrieval overlaps and novelty distribution. *Journal of the American Society for Information Science*, 40(2), 110-114.

Merton, R. K. (1942): Science and technology in a democratic order. *Journal of Legal and Political Sociology*, 1, 115-126. (Reprinted several times in slightly revised versions with different titles, e.g., as "Science and democratic social structure." in Merton, R. K. (1967): *Social theory and social structure, 2nd revised edition*. New York: Free Press, p. 550-561, as "The institutional imperatives of science." in Barnes, B. S. ed. (1972): *The sociology of science*. Harmondsworth, Penguin, p. 65-79, and as "The normative structure of science." in Storer, N. W. ed. (1973): *The sociology of science : theoretical and empirical investigations*. Chicago: University of Chicago Press, p. 267-278.)

Merton, R. K. (1988): The Matthew effect in science II : cumulative advantage and the symbolism of intellectual property. *ISIS*, 79(299), 606-623. [http://garfield.library.upenn.edu/merton/matthewii.pdf, visited 7-6-2003]

Michie, D. (1974): *On machine intelligence*. Edinburgh: Edinburgh University Press.

Moed, H. F., De Bruin, R. E. and Van Leeuwen, Th. N. (1995): New bibliometric tools for the assessment of national research performance : database description, overview of indicators, and first applications. *Scientometrics*, 33, 381-422.

Moravcsik, M. J. and Murugesan, P. (1975): Some results on the function and quality of citations. *Social Studies of Science*, 5(1), 86-92.

Morman, E. T. (1981): Citation analysis and the current debate over quantitative methods in the social studies of science. *Society for the Social Studies of Science Newsletter*, 5(3), 7-13. (Referred in Cronin, 1984. The original paper could not be obtained)

Mählck, P. and Persson, O. (2000): Building data for network analysis. *Scientometrics*, 49(1), 81-91.

Otter, M. and Johnson, H. (2000): Lost in hyperspace : metrics and mental models. *Interacting with computers*, 13, 1-40.

Over, P. (1997): TREC-5 interactive track report. In: Voorhees, E. M. and Harman, D. K. eds. *The fifth text retrieval conference (TREC-5)*. [Gaithersburg, MD]: National Institute of Standards and Technology, p. 29-56. (NIST Special Publication ; 500-238) [http://trec.nist.gov/pubs/trec5/papers/trackreport.ps.gz, visited 31-10-2003]

Pao, M. L. (1984): Semantic and pragmatic retrieval. In: Flood, B., Witiak, J. and Hogan, T. H. eds. *Challenges to an information society : proceedings of the 47th ASIS annual meeting : Philadelphia, Pennsylvania, October 21-25, 1984*. New York: American Society for Information Science, p. 134-136.

Pao, M. L. (1993): Term and citation retrieval - a field-study. *Information Processing & Management*, 29(1), 95-112.

Pao, M. L. and Fu, T. T. W. (1985): Titles retrieved from MEDLINE & from citation relations. In: Parkhurst, C. A. ed. *ASIS '85 : proceedings of the 48th ASIS annual meeting, Las Vegas, Nevada, October 20-24, 1985*. Medford, N.J.: Learned Information, p. 120-123.

Pao, M. L. and Worthen, D. B. (1989): Retrieval effectiveness by semantic and citation searching. *Journal of the American Society for Information Science*, 40(4), 226-235.

Pirkola, A., Keskustalo, H., Leppänen, E., Känsälä, A.-P. and Järvelin, K. (2002): Targeted s-gram matching : a novel n-gram matching technique for cross- and monolingual word form variants. *Information Research*, 7(2), -paper no. 126. [http://informationr.net/ir/7-2/paper126.html, visited 23-8-2003]

Pollack, S. M. (1968): Measures for the comparison of information retrieval systems. *American Documentation*, 19(4), 387-397.

Porter, M. (1980): An algorithm for suffix stripping. *Program*, 14(3), 130-137.

Price, D. J. d. S. (1970): Citation measures of hard science, soft science, technology and nonscience. In: Nelson, C. E. and Pollock, D. K. eds. *Communication among scientists, and engineers*. Lexington, Mass.: Heath Lexington Books, p. 3-22. (Studies in social and economic process)

Qin, J. (2000): Semantic similarities between a keyword database and a controlled vocabulary database : an investigation in the antibiotic resistance literature. *Journal of the American Society for Information Science*, 51(3), 166-180.

van Raan, A. F. J. (1998): In matters of quantitative studies of science the fault of theorists is offering too little and asking too much. *Scientometrics*, 43(1), 129-139.

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.

Rehfeld, J. F. (1978a): Immunochemical studies on cholecystokinin : 1 : development of sequence-specific radioimmunoassays for porcine triacontatriapeptide cholecystokinin. *Journal of Biological Chemistry* , 253(11), 4016-4021. [http://www.jbc.org/cgi/reprint/253/11/4016.pdf, visited 10-8-2003a]

Rehfeld, J. F. (1978b): Immunochemical studies on cholecystokinin : 2 : distribution and molecular heterogeneity in central nervous-system and small-intestine of man and hog. *Journal of Biological Chemistry*, 253(11), 4022-4030. [http://www.jbc.org/cgi/reprint/253/11/4022.pdf, visited 10-8-2003b]

Rehfeld, J. F. (1981): 4 basic characteristics of the gastrin-cholecystokinin system. *American Journal of Physiology*, 240(4), G255-G266.

van Rijsbergen, C. J. (1979): *Information retrieval*. London: Butterworths. ix, 208 p. [http://www.dcs.gla.ac.uk/Keith/Preface.html ; http://www.dcs.gla.ac.uk/~iain/keith/, visited 3-9-2003]

Rivlin, E., Botafogo, R. A. and Shneiderman, B. (1994): Navigating in hyperspace: designing a structure-based toolbox. *Communications of the ACM*, 37(2), 87-96.

Robertson, S. E. and Sparck Jones, K. (1976): Relevance weighting of search terms. *Journal of the American Society for Information Science*, 27, 129-146.

Robertson, S. E. (1981): The methodology of information retrieval experiment (Chapter 2). In: Sparck Jones, K. ed. *Information retrieval experiment*. London: Butterworths, p. 9-31.

Rocchio, J. J. (1971): Relevance feedback in information retrieval. In: Salton, G. ed. *The SMART retrieval system : experiments in automatic document processing*. Englewood Cliffs, NJ: Prentice Hall, p. 313-323. (Prentice-Hall series in automatic computation)

Roelleke, T., Lalmas, M., Kazai, G., Ruthven, I. and Quicker, S. (2002): The accessibility dimension for structured document retrieval. In: Crestani, F., Girolami, M. and van Rijsbergen, C. J. eds. *Advances in information retrieval : 24th NCS-IRSG European colloquium on ir research : Glasgow, UK, March 25-27, 2002 : proceedings*. Berlin: Springer-Verlag, p. 284-302. (Lecture Notes in Computer Science ; LNCS 2291)

Rousseau, R. (1998): Venn, Carroll, Karnaugh en Edwards [Venn, Carroll and Edwards]. *Wiskunde & Onderwijs*, 24(95), 233-241. (In Dutch)

Ruskey, F. (2001): A Survey of Venn Diagrams. *The Electronic Journal of Combinatorics*, (Dynamic survey, last update March 2001) [http://www.combinatorics.org/Surveys/ds5/VennEJC.html, visited 2-8-2003]

Ruthven, I. (2001): *Retrieval through explanation: an abductive inference approach to relevance feedback*. Glasgow: Department of Computing Science, University of Glasgow. 555 p. [http://www.cs.strath.ac.uk/~ir/papers/thesis.pdf, visited 11-11-2003]

Ruthven, I. and Lalmas, M. (1998): Representing and retrieving structured documents using the Dempster-Shafer theory of evidence : modelling and evaluation. *Journal of Documentation*, 54(5), 529-565.

Salton, G. (1962): Some experiments in the generation of words and document associations. In: *Proceedings of the AFIPS fall joint computer conference*. Philadelphia: Spartan Books, p. 234-250.

Salton, G. (1963): Associative document retrieval techniques using bibliographic information. *Journal of the Association for Computing Machinery*, 10(4), 440-457.

Salton, G. (1971): Automatic indexing using bibliographic citations. *Journal of Documentation*, 27(2), 98-110.

Salton, G. and McGill, M. J. (1983): *Introduction to modern information retrieval*. New York: McGraw-Hill Book Company. 448 p. (McGraw-Hill Computer Science Series) ISBN: 0-07-054484-0.

Salton, G. and Buckley, C. (1988): Term-weighting approaches in automatic text retrieval. *Information Processing & Management*, 24(5), 513-523.

Salton, G. and Zhang, Y. (1986): Enhancement of text representations using related document titles. *Information Processing & Management*, 22(5), 385-394.

Salton, G., Allan, J. and Buckley, C. (1993): Approaches to passage retrieval in full text information systems. In: Korfhage, R., Rasmussen, E. and Willett, P. eds. *SIGIR'93 : proceedings of the 16th annual international ACM SIGIR conference on research and development in information retrieval*. New York: ACM Press, p. 49-58. [http://www.ai.mit.edu/people/jimmylin/papers/Salton93.pdf (preprint), visited 24-9-2003]

Saracevic, T. (1969): Comparative effect of titles, abstracts, and full texts on relevance judgement. In: North, J. B. ed. *ASIS '69 : proceedings of the 32nd ASIS annual meeting, San Francisco, California, October, 1-4, 1969*. Westport, Conn.: Greenwood Press, p. 293-299.

Saracevic, T. (1996): Relevance reconsidered '96. In: Ingwersen, P. and Pors, N. O. eds. *Proceedings : CoLIS2 : second international conference on conceptions of library and information science : integration in perspective : October 1-16, 1996 : the Royal School of Librarianship*. København: The Royal School of Librarianship, p. 200-218.

Saracevic, T., Kantor, P., Chamis, A. Y. and Trivison, D. (1988): A study of information seeking and retrieving : [part] I : background and methodology. *Journal of the American Society for Information Science*, 39(3), 161-176.

Saracevic, T. and Kantor, P. (1988a): A study of information seeking and retrieving : [part] II : user, questions, and effectiveness. *Journal of the American Society for Information Science*, 39(3), 177-196.

Saracevic, T. and Kantor, P. (1988b): A study of information seeking and retrieving : [part] III : searchers, searches, and overlap. *Journal of the American Society for Information Science*, 39(3), 197-216.

Saracevic, T. and Su, L. (1989): Modelling and measuring user-intermediary-computer interaction online searching : design of study. In: Katzer, J. and Newby, G. B. eds. *ASIS '89 : proceedings of the 52nd ASIS annual meeting, Washington, DC October 30-November 2, 1989*. Medford, N.J.: Learned Information, p. 75-80.

Saracevic, T., Mokros, H. and Su, L. (1990): Nature of interaction between users and intermediaries in online searching : a qualitative analysis. In: Henderson, D. ed. *ASIS '90 : proceedings of the 53rd ASIS annual meeting, Toronto, Ontario November 4-8, 1990*. Medford, N.J.: Learned Information, p. 47-54. (ASIS ; 27)

Saracevic, T., Mokros, H., Su, L. and Spink, A. (1991): Interaction between users and intermediaries in online searching. In: Williams, M. E. ed. *National online meeting : proceedings 1991*. Medford, N.J.: Learned Information, p. 329-342.

Schamber, L., Eisenberg, M. B. and Nilan, M. S. (1990): A re-examination of relevance : towards a dynamic, situational definition. *Information Processing & Management*, 26(6), 755-776.

Schatz, B., Mischo, W. H., Cole, T. W., Hardin, J. B. and Bishop, A. P. (1996): Federating diverse collections of scientific literature. *IEEE Computer*, 29(5), 28-36.

Schütz, T. (1998): *Retrieval von komplexen Objekten am Beispiel von SGML-Dokumenten [Retrieval of complex objects, considering SGML as example]*. [Dortmund]: Universität Dortmund, Fachbereich Informatik. [iv], 98 p. (Unpublished masters thesis - in German) [http://ls6-www.informatik.uni-dortmund.de/~roelleke/DPAs/schuetz.ps.gz, visited 13-8-2003]

Seglen, P. O. (1997): Why the impact factor of journals should not be used for evaluating research. *British Medical Journal*, 314(7079), 498-502.

Shaw, W. M. Jr. (1990a): Subject indexing and citation indexing : part 1 : clustering structure in the cystic fibrosis collection. *Information Processing & Management*, 26(6), 693-703.

Shaw, W. M. Jr. (1990b): Subject indexing and citation indexing : part 2 : an evaluation and comparison. *Information Processing & Management*, 26(6), 705-718.

Shaw, W. M. Jr. (1991a): Subject and citation indexing : part I : The clustering structure of composite representations in the cystic fibrosis document collection. *Journal of the American Society for Information Science*, 42(9), 669-675.

Shaw, W. M. Jr. (1991b): Subject and citation indexing : part II : the optimal cluster-based retrieval performance of composite representations. *Journal of the American Society for Information Science*, 42(9), 676-684.

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.

Siegel, S. and Castellan, N. J. Jr. (1988): *Nonparametric statistics for the behavioral sciences*. New York: McGraw-Hill Book Company. xxiii, 399 p. (McGraw-Hill international editions - statistical series) ISBN: 0-07-057357-3.

Small, H. (1973): Co-citation in the scientific literature : a new measure of the relationship between two documents. *Journal of the American Society for Information Science*, 24, 265-269.

Small, H. G. (1978): Cited documents as concept symbols. *Social Studies of Science*, 8, 327-340.

Small, H. (1982): Citation context analysis. In: Dervin, B. J. and Voigt, M. J. eds. *Progress in Communications Sciences*. Norwood, N. J.: Ablex, vol. 3, p. 287-310.

Small, H. (1987): Report on citation analysis research at ISI. In: *Science literacy, policy, evaluation, and other essays : essays of an information scientist: 1988, vol. 11*. [Philadelphia]: [ISI Press], p. 381-385. [Is contained in: http://www.garfield.library.upenn.edu/essays/v11p160y1988.pdf, visited 15-6-2003]

Small, H. G. and Griffith, B. C. (1974): The structure of scientific literatures I : identifying and graphing specialities. *Science Studies*, 4, 17-40.

Sormunen, E. (2002): Liberal relevance criteria of TREC : counting on negligible documents? In: Beaulieu, M., Baeza-Yates, R., Myaeng, S. H. and Järvelin, K. eds. *Proceedings of SIGIR 2002 : the twenty-fifth annual international ACM SIGIR conference on research and development in information retrieval, August 11-15, 2002, Tampere, Finland*. New York: ACM Press, p. 324-330.

Sparck Jones, K. (1974): Automatic indexing. *Journal of Documentation*, 30(4), 393-432.

Sparck Jones, K. (1998): Summary performance comparisons TREC-2 through TREC-7. In: Voorhees, E. M. and Harman, D. K. eds. *The Seventh Text REtrieval Conference (TREC-7)*. [Gaithersburg, MD]: National Institute of Standards and Technology, p. B-1-B-6. (NIST Special Publication ; 500-242) [http://trec.nist.gov/pubs/trec7/papers/sparckjones.pdf.gz, visited 17-8-2003]

Sparck Jones, K. and van Rijsbergen, C. J. (1976): Information retrieval test collections. *Journal of Documentation*, 32(1), 59-75.

Spink, A. (1996): Multiple search sessions model of end-user behavior : an exploratory study. *Journal of the American Society for Information Science*, 47(8), 603-609.

Spink, A. (1997): Study of interactive feedback during mediated information retrieval. *Journal of the American Society for Information Science*, 48(5), 382-394.

Spärck Jones, K. (2003): Document retrieval : shallow data, deep theories; historical reflections, potential directions. In: Sebastiani, F. ed. *Advances in information retrieval : 25th European conference on IR research, ECIR 2003, Pisa, Italy, April 14-16, 2003 : proceedings*. Berlin: Springer, p. 1-11. (Lecture Notes in Computer Science ; 2633)

Su, L. T. (1992): Evaluation measure for interactive information retrieval. *Information Processing & Management*, 28, 503-516.

Su, L. T. (1994): The relevance of recall and precision in user evaluation. *Journal of the American Society for Information Science*, 45(3), 207-217.

Sumner, R. G. Jr. (1995): Utilizing the age of references to control the exhaustivity of the reference representation in information retrieval. *Information Processing & Management*, 31(1), 29-45.

Swales, J. M. (1990): Research articles in English (Chapter 7). In: Long, M. H. and Richards, J. C. eds. *Genre analysis : English in academic and research settings*. Cambridge: Cambridge University Press, p. 110-176. (Cambridge Applied Linguistics)

Taylor, R. S. (1968): Question negotiation and information seeking in libraries. *College and Research Libraries*, 29, 178-194.

Thellwall, M. (2003): What is this link doing here : beginning a fine-grained process of identifying reasons for academic hyperlink creation. *Information Research*, 8(2), paper no. 151. [http://informationr.net/ir/8-3/paper151.html, visited 8-11-2003]

Thorne, F. C. (1977): The citation index : another case of spurious validity. *Journal of Clinical Psychology*, 33(3), 1157-1161.

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]

Turtle, H. and Croft, W. B. (1990): Inference networks for document retrieval. In: Vidick, J.-L. 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. and Croft, W. B. (1991): Evaluation of an inference network-based retrieval model. *ACM Transactions on Information Systems*, 9(3), 187-222.

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]

Venn, J. (1880): On the diagrammatic and mechanical representation of propositions and reasonings. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 9, 1-18.

Vladutz, G. and Cook, J. (1984): Bibliographic coupling and subject relatedness. In: Flood, B., Witiak, J. and Hogan, T. H. eds. *Challenges to an information society : proceedings of the 47th ASIS annual meeting : Philadelphia, Pennsylvania, October 21-25, 1984*. New York: American Society for Information Science, p. 204-207.

Voorhees, E. M. (2001): Evaluation by highly relevant documents. In: Croft, W. B., Harper, D. J., Kraft, D. H. and Zobel, J. eds. *Proceedings of the 24th annual international ACM SIGIR conference on*

*research and development in information retrieval : New Orleans, Louisiana, USA, September 9-13, 2001*. New York: ACM Press, p. 74-82.

Voorhees, E. M. and Harman, D. K. eds. (2002): *The tenth text retrieval conference (TREC 2001)*. [Gaithersburg, MD]: National Institute of Standards and Technology. 976 p. (NIST Special Publication ; 500-250) ISBN: 0-16-067409-3. [http://trec.nist.gov/pubs/trec10/t10_proceedings.html, visited 9-8-2003]

Weinstock, M. (1971): Citation Indexes. In: Allen, K. ed. *Encyclopeadia of Library and Information Science, vol. 5*. New York: Marcel Decker Inc., p. 16-40. (Reprinted in: *Essays of an information scientist: 1962-1973, vol. 1*, 1977, p. 188-195, 198-205, 207-216) [http://www.garfield.library.upenn.edu/essays/V1p188y1962-73.pdf, visited 8-7-2003]

White, H. D. and Griffith, B. C. (1981): Author cocitation : a literature measure of intellectual structure. *Journal of the American Society for Information Science*, 32, 163-171.

White, H. D. and McCain, K. W. (1998): Visualizing a discipline : an author co-citation analysis of information science, 1972-1995. *Journal of the American Society for Information Science*, 49(4), 327-355.

Wilson, T. D. (1984): The cognitive approach to information seeking behaviour and information use. *Journal of Documentation*, 55(3), 249-270.

Wormell, I. (1981): SAP - a new way to produce subject descriptions of books. *Journal of Information Science : Principles and Practice*, 3, 39-43.

Wormell, I. (1985): *Subject Access Project : SAP : improved subject retrieval for monographic publications*. Lund: Lund University. 141 p.

Wouters, P. (1999): Beyond the Holy Grail : from citation theory to indicator theories. *Scientometrics*, 44(3), 561-580.

Wu, M.-M. and Liu, Y.-H. Intermediary's information seeking, inquiring minds, and elicitation styles. Journal of the American Society for Information Science, 54(12), 1117-1133.

# List of abbreviations

| | |
|---|---|
| A&HCI | Arts and Humanities Citation Index |
| AB | Abstract |
| ABS | Abstract (Document representation) |
| AKW | Author keywords (Document representation) |
| ASK | Anomalous State of Knowledge |
| ATL | Article title (Document representation) |
| AvgP | Average Precision |
| CA | Cited Author |
| CAS topics | Content And Structure topics |
| CCV_step2 | Citation Cut-off Value for Step 2. Variable in the best match boomerang effect that determines the number of seed documents used. Can take three values Low, medium and High, corresponding to 25, 50 or 75 % of the available citations. |
| CG | Cumulated Gain |
| CLEF | Cross Language Evaluation Forum |
| CO topics | Content Only topics |
| CON | Conclusions (Document representation) |
| CR | Cited Reference |
| CTI | Cited titles (Document representation) |
| CY | Cited Year |
| DCG | Discounted Cumulated Gain |
| DCV | Document Cut-off Value |
| DCV_step1 | Document Cut-off Value for Step 1. Variable in the best match boomerang effect that determines the size of the sets of retrieved documents in Step 1. |
| DE | Descriptor |
| DE | Descriptors (Document representation) |
| DTD | Document Type Definition |
| ESL | Expected Search length |
| ff | Variable in the best match boomerang effect. Denotes that a flat citation index was used in Step 1 and Step 3. |
| FGC | Figure captions (Document representation) |

| | |
|---|---|
| fx | Variable in the best match boomerang effect. Denotes that a flat citation index was used in Step 1 and an extended in Step 3. |
| gen_inex | Generalized INEX quantification function defined by the INEX organisers. |
| gen_whole | Generalized quantification function adapted to whole documents for the dissertation work. |
| gP | generalized Precision |
| gR | generalized Recall |
| HTML | Hypertext Mark-up Language |
| $I^3R$ | Intelligent Intermediary for Information Retrieval |
| ID | Identifiers (Document representation) |
| IDF | Inverse Document Frequency |
| IEEE | Institute for Electrical and Electronics Engineers |
| IIR | Interactive Information Retrieval |
| INEX | The Initiative for the Evaluation of XML retrieval |
| INT | Introductions (Document representation) |
| IR | Information Retrieval |
| ISI | Institute for Scientific Information |
| IT | Information Technology |
| KWIC | KeyWords In Context |
| MB | Mega Byte |
| nCG | normalised Cumulated Gain |
| nDCG | normalised Discounted Cumulated Gain |
| OL(n) | E.g., OL3, Overlap Level |
| PDF | Portable Document Format |
| P-R curves | Precision Recall curves |
| PS | PostScript |
| R&D | Research and Development |
| RHL | Ranked Half Life |
| RR | Relative Relevance |
| SAP | Subject Access Project |
| SCI | Science Citation Index |
| SGML | Standard Generalized Mark-up Language |
| SSCI | Social Science Citation Index |

| | |
|---|---|
| str_inex | Strict INEX quantification function defined by the INEX organisers. |
| str_whole | Strict quantification function adapted to whole documents for the dissertation work. |
| TAPIR | Text Access Potentials for Interactive Information Retrieval |
| TBC | Table captions (Document representation) |
| TF | Term Frequency |
| TI | Title |
| TREC | Text REtrieval Conference |
| URID | Unique Reference IDs |
| WWW | World Wide Web |
| xf | Variable in the best match boomerang effect. Denotes that an extended citation index was used in Step 1 and a flat in Step 3. |
| XML | eXtensible Mark-up Language |
| xx | Variable in the best match boomerang effect. Denotes that a flat citation index was used in Step 1 and Step 3. |

# Appendices

# Appendix 1: The three work tasks used in the pre-experiment

The work tasks consisted of a verbal formulation and the actual Boolean search statement used.

**Simulated work task 1:**
As a researcher your main research interest is osteoporosis (brittleness of the bones). An important question in your research is the problem of how to detect patients who have bone fractures, for which ultrasound measurements are commonly applied as a diagnostic tool. You are interested in identifying evidence in the literature as to how well quantitative ultrasound (QUS) separate those patients who suffer from bone fractures from those that do not, and to what extent QUS can be used to predict which patients will suffer from bone fractures in the future.

(bua OR sos OR qus) AND fracture*

**Simulated work task 2:**
As a researcher your main research interest is osteoporosis (brittleness of the bones). In order to prevent bone fractures in patients suffering from osteoporosis it is important to maximise and maintain bone strength. The bone mineral content (BMC) and bone mineral density (BMD) are important indicators of skeletal strength. Earlier studies have indicated that that BMC and BMD are influenced by physical activity, and you are interested in obtaining evidence from the literature of the effect of physical exercise and stimulation on bone strength as well as muscle strength.

(BMD OR BMC OR mineral density OR mineral content OR bone density OR bone mass) AND bone AND (musc* strength OR bone strength) AND (exerci* OR stimulat* OR train* OR physical activit*)

**Simulated work task 3:**
As a researcher your main research interest is osteoporosis (brittleness of the bones). Earlier research has been shown that osteoporosis is influenced by a variety of factors, such as physical activity, age and sex of patients, whether women patients are pre- or postmenopausal etc. You are interested in finding evidence from the literature that can indicate how genetic factors affect osteoporosis, and if there is a risk of osteoporosis to be a hereditary disease. You give special attention to studies that compare monovular twins with biovular twins.

(BMD OR bone mass OR BMC OR osteoporo* OR fracture* OR BUA OR SOS OR QUS) AND (genetic OR heredit* OR polymor* OR mutat*) AND (twin* OR gemel*)

# Appendix 2: Statistics on the INEX corpus

Distribution of the INEX corpus on journals with statistics for the number of articles, the file size, and the Journal Impact Factor from Journal Citation Reports 2002.

| id | Journal title | Years | Size (MB) | no of articles | JIF |
|---|---|---|---|---|---|
| an | IEEE Annals of the History of Computing | 1995-2001 | 13.2 | 316 | 0.222 |
| cg | IEEE Computer Graphics and Applications | 1995-2001 | 19.1 | 680 | 1.193 |
| co | Computer | 1995-2001 | 40.4 | 1,902 | 1.031 |
| cs | IEEE Computational Science & Engineering | 1995-1998 | | | - |
| | Computing in Science & Engineering | 1999-2001 | 14.6 | 571 | 0.795 |
| dt | IEEE Design & Test of Computers | 1995-2001 | 13.6 | 539 | 1.145 |
| ex | IEEE Expert | 1995-1997 | 20.3 | 702 | - |
| | IEEE Intelligent Systems | 1998-2001 | | | 1.905 |
| ic | IEEE Internet Computing | 1997-2001 | 12.2 | 547 | 1.024 |
| it | IT Professional | 1999-2001 | 4.7 | 249 | - |
| mi | IEEE Micro | 1995-2001 | 15.8 | 604 | 1.065 |
| mu | IEEE Multimedia | 1995-2001 | 11.3 | 465 | 0.481 |
| pd | IEEE Parallel & Distributed Technology | 1995-1996 | | | - |
| | IEEE Concurrency | 1997-2000 | 10.7 | 363 | 0.515 |
| so | IEEE Software | 1995-2001 | 20.9 | 936 | 1.068 |
| tc | IEEE Transactions on Computers | 1995-2002 | 66.1 | 1,042 | 1.484 |
| td | IEEE Transactions on Parallel & Distributed Systems | 1995-2002 | 58.8 | 765 | 0.819 |
| tg | IEEE Transactions on Visualization & Computer Graphics | 1995-2002 | 15.2 | 225 | 1.741 |
| tk | IEEE Transactions on Knowledge and Data Engineering | 1995-2002 | 48.1 | 585 | 1.055 |
| tp | IEEE Transactions on Pattern Analysis and Machine Intelligence | 1995-2002 | 62.9 | 1,046 | 2.923 |
| ts | IEEE Transactions on Software Engineering | 1995-2002 | 46.1 | 570 | 1.170 |
| | | | 494 | 12,107 | |

# Appendix 3: Sample references in XML mark-up

Sample references formatted in the original XML mark-up from the article:

Liu, J.C.L., Du, D.H.C., Shim, S.S.Y., Hsieh, J., and Lin, M. (1999): Design and Evaluation of a Generic Software Architecture for On-Demand Video Servers. *IEEE Transactions on Knowledge and Data Engineering*, 11(3), 406-424. (file name: k0406.xml)

The following tags are used in the formatting of the content:

| | |
|---|---|
| <bb> | Delimits a reference. Includes an ID that is unique for the referring article. |
| <au> | Name of cited author, with first name (<fnm>) and sir name (<snm>). |
| <atl> | The title of the cited paper or article. |
| <ti> | The title of the cited carrier, e.g., a journal, a conference proceeding or a book. |
| <volno> | Cited volume number. |
| <issno> | Cited issue number. |
| <pp> | Cited page numbers. |
| <pdt> | Cited publication details, with month (<mo>) and year (<yr>). |
| <obi> | A tag used or miscellaneous purposes, e.g., to insert words or to mark elements that are not explicitly part of the DTD. |

The portions of underlined text show the content that will be printed as the result of an XSL transformation of a given tag. In this process new tags or delimiters can be inserted between the content of the tags. Reference 1, 5, 16, 38 and 40 are shown as examples of references to different types of publications: Reference *bibk04061* is to a journal article, *bibk04065* and *bibk040616* to conference papers, and *bibk040638* and *bibk040640* to book chapters.

```
<bb id="bibk04061">
     <au><fnm>D.</fnm><snm>Anderson</snm></au>
     <au><fnm>Y.</fnm><snm>Osawa</snm></au><obi>and</obi>
     <au><fnm>R.</fnm><snm>Govindan</snm></au>
     <atl>"A File System for Continuous Media,"</atl>
     <ti>ACM Trans. Computer Systems,</ti>
     <obi><volno>vol. 10,</volno><issno>no. 4,</issno></obi>
     <pp>pp. 311-337,</pp>
     <pdt><mo>Nov.</mo><yr>1992.</yr></pdt>
</bb>
```

```
<bb id="bibk04065">
     <au><fnm>T.</fnm><snm>Chua</snm></au>
     <au><fnm>J.</fnm><snm>Li</snm></au>
     <au><fnm>B.</fnm><snm>Ooi</snm></au><obi>and</obi>
     <au><fnm>K.</fnm><snm>Tan</snm></au>
     <atl>"Disk Striping Strategies for Large Video-On-Demand Servers,"</atl>
     <ti>Proc. ACM Multimedia '96 Conf.,</ti>
     <pp>pp. 297-306,</pp>
     <pdt><mo>Nov.</mo><yr>1996.</yr></pdt>
</bb>

<bb id="bibk040616">
     <au><fnm>J.</fnm><snm>Hsieh</snm></au>
     <au><fnm>D.</fnm><snm>Du</snm></au>
     <au><fnm>J.</fnm><snm>MacDonald</snm></au>
     <au><fnm>J.</fnm><snm>Thomas</snm></au>
     <au><fnm>J.</fnm><snm>Pugaczewski</snm></au>
     <au><fnm>J.</fnm><snm>Kays</snm></au><obi>and</obi>
     <au><fnm>M.</fnm><snm>Wiklund</snm></au>
     <atl>"Experimental Study of Extended HIPPI Connections Over ATM
          Networks,"</atl>
     <ti>Proc. 15th Ann. Joint Conf. IEEE Computer and Comm. Societies (IEEE
          Infocom '96),</ti>
     <loc><cty>San Francisco, Calif.,</cty></loc>
     <pp>pp. 1,261-1,268,</pp>
     <pdt><mo>Mar.</mo><yr>1996.</yr></pdt>
</bb>

<bb id="bibk040638">
     <au><fnm>A.</fnm><snm>Silberschatz</snm></au>
     <au><fnm>J.</fnm><snm>Peterson</snm></au><obi>and</obi>
     <au><fnm>P.</fnm><snm>Galvin</snm></au>
     <ti>Operating System Concepts,</ti>
     <obi>third ed.,</obi><obi>ch. 4,</obi><obi>Addison-Wesley,</obi>
          <pdt><yr>1991.</yr></pdt>
</bb>

<bb id="bibk040640">
     <au><fnm>R.</fnm><snm>Stevens</snm></au>
     <ti>Unix Network Programming,</ti>
     <obi>fourth ed.,</obi><obi>ch. 3,</obi><obi>Prentice Hall,</obi>
     <pdt><yr>1990.</yr></pdt>
</bb>
```

# Appendix 4: Identification of citations to the same document

In order to facilitate the discovery of citations to the same document it was attempted to reduce the errors and inconsistencies in the cited paper titles. This was done with the edit distance algorithm in the like program, which was kindly provided by the Department of Information Studies, University of Tampere, Finland.

The calculation of edit distances can be computationally expensive on long strings since every character of the two strings has to be taken into consideration. Since some strings were very long they were cut down to the first 100 characters in order to reduce processing time (4.3% of the cited paper titles were more than 100 characters long). Even with these reductions matching all the 118,191 cited paper titles (<ti>) against themselves will result in a very large amount of matches (more than 13 billion matches). Therefore it was chosen to do an ad hoc partitioning of the cited paper titles by cited year in order to reduce the number of matches. The primary concern in the partitioning was to make groups of strings of a manageable size. After initial tests it was found that groups of approximately 6,000 to 10,000 strings could be processed within the time available. This resulted in the partitioning shown in the table below, where some groups consist of single cited years and others were ranges of years[72]. A small group of 1926 cited paper titles did not have a cited year. These were omitted from the matching because other matching titles were probably to be found in the other groups rather than within this group.

With the partitions the number of matches was reduced to 928.4 million matches, or 6.7% of the amount of matches that should have been carried out of all strings were matched against all strings. As this was still a large amount of matches it was chosen to run the groups on separate servers in parallel in order to speed up the processing. A small part of the matches were run on the server in the IR laboratory at the Department of Information Studies, University of Tampere, Finland, and the remaining parts were run in the IT laboratory at the Department for Information Studies at the Royal School of Library and Information Science in Copenhagen. In Copenhagen it was attempted to use the source code of the like program to make a full version for Linux complete with the index structure and all the approximate string matching algorithms, but insufficient time was available to complete this. Instead the edit distance component was taken out and a implemented in a simple version without the index structure, which ran in a linear mode instead. This was not nearly as efficient as with the index structure, but could run

[72] The table also provides a rough indication of the age distribution of the citations (excluding references to monographs and those without a cited year) when taking into to consideration that the articles in the corpus as a whole was published in 1995-2002.

on all available machines. The run-time of to do 928.4 million edit distance matches was approximately one full week on ten 1400 MHz machines.

| Cited years | No. of cited paper titles (<atl>) | No. of edit distance calculations |
|---|---|---|
| -1959 | 1,296 | 1,679,616 |
| 1960-1978 | 6,033 | 36,397,089 |
| 1979-1983 | 6,051 | 36,614,601 |
| 1984-1986 | 7,171 | 51,423,241 |
| 1987-1988 | 8,171 | 66,765,241 |
| 1989 | 5,033 | 25,331,089 |
| 1990 | 6,745 | 45,495,025 |
| 1991 | 7,502 | 56,280,004 |
| 1992 | 8,704 | 75,759,616 |
| 1993 | 10,330 | 106,708,900 |
| 1994 | 10,673 | 113,912,929 |
| 1995 | 10,399 | 108,139,201 |
| 1996 | 8,838 | 78,110,244 |
| 1997 | 7,403 | 54,804,409 |
| 1998 | 5,842 | 34,128,964 |
| 1999-2002 | 6,074 | 36,893,476 |
| NA | (1,926) | - |
| Total (un-partitioned) | 118,191 | 13,969,112,481 |
| Total (partitioned) | 116,265 | 928,443,645 |

As described in Section 6.2.4 the cited paper titles were changed to small letters, any punctuation removed, and non-English characters changed to blanks in order to facilitate a better match, and the longest strings reduced to 100 characters. In each group every of these strings were matched against all other strings using the edit distance. The output from like is a list of strings that have the smallest edit distance to the query string up to a given threshold. After initial tests it was found that a threshold of 30 would retrieve all reasonable variants of the strings even for the longest strings of 100 characters. This output was used to calculate a weighted edit distance relative to the length of the query string. The weighted edit distance was ultimately used to replace the worst of the identified strings with the best ones through a number of steps, which are outlined below. No test corpus was available to measure to performance of these operations. Instead the results were inspected and a threshold for the weighted edit distance chosen that seemed to give the best results. A threshold of 0.25 was chosen, that is if for instance an absolute edit distance of 13 was found to a query string with 50 characters, the answer string was not included (13/50 = 0.26), whereas an absolute edit distance of 12 would include the string in the replacement pool. (12/50 = 0.24). This resulted in a correction of many of the errors and inconsistencies in the cited paper titles. These were then sorted by cited titles within each group and a unique identifier assigned to each so that citations to the article could be found across the corpus. Before the corrections there were 85,707 unique strings among the 116,265 cited paper titles with a cited year. After the corrections there were 70,634 unique strings, corresponding to a reduction of 17.6%.

**Steps in the use of like output to identify identical citations**

1. Calculate weighted edit distance values in relation to the length of the query string for each of the answer strings, and remove any answer strings with values above the threshold of 0.25.
2. Order all cited paper titles in the group by the "best" query strings (defined as the strings that have the highest number of matching answer strings with the lowest average weighted edit distances).
3. Replace each answer string with its query string beginning with answers strings for the "worst" query strings and ending with the answer strings for the "best" query strings. Thereby the strings that had the most matches with the lowest weighted edit distances will be grouped together by the most frequently occurring string in the group.

# Appendix 5: Example of a citation query

Below is shown a weighted list of citations from Step 2 in the best match boomerang effect formatted as a weighted sum query to be submitted in Step 3. This is the list for topic 39 (see Figure 6.2 and Figure 6.7) with DCV_step1 = 32, citations extracted from the flat citation index for Step 2, CCV_step2 = 0.50. The URIDs are unique reference IDs as described in Section 6.2.4 on the generation of a citation code.

```
#q39 = #WSUM(1 1238.55 URID041197 1074.78 URID061534
1053.66 URID043306 1024.69 URID036037 900.054 URID037126
775.504 URID050846 774.951 URID057053 710.601 URID057017
710.515 URID062777 645.914 URID047061 579.066 URID056216
563.284 URID043619 560.457 URID041914 557.323 URID048752
551.271 URID033622 544.039 URID046439 532.951 URID035106
530.606 URID025020 512.218 URID049657 505.280 URID027019
489.943 URID039919 489.208 URID050202 484.435 URID054611
479.822 URID052967 456.653 URID051978 448.117 URID051243
420.653 URID028505 416.489 URID011553 406.368 URID039175
397.482 URID057067 394.510 URID031499 391.233 URID041233
391.233 URID045027 391.233 URID058899 389.838 URID049931
381.372 URID049733 370.119 URID033855 363.313 URID070973
361.252 URID058591 359.093 URID034689 357.182 URID006191
354.660 URID038829 353.801 URID046746 350.154 URID068927
343.566 URID037419 340.174 URID050394 334.307 URID028459
333.199 URID032554 331.905 URID039922 321.275 URID036730
301.367 URID039551 298.696 URID056770 293.425 URID045449
293.425 URID062770 291.940 URID070946 290.636 URID048208
289.100 URID033119 287.272 URID053884 286.371 URID032061
284.206 URID059620 272.485 URID064507 271.666 URID053409
265.995 URID028576 265.679 URID048915 262.969 URID041052
262.969 URID041368 262.969 URID055647 259.503 URID056459
255.831 URID061748 254.795 URID030413 253.917 URID047422
251.727 URID042935 247.620 URID054648 238.957 URID045469
238.957 URID064757 236.165 URID041338 228.327 URID043279
227.685 URID053116 227.071 URID068476 227.071 URID071989
223.025 URID042856 221.110 URID035411 215.262 URID030716
215.262 URID042035 213.154 URID034654 213.154 URID066439
212.332 URID045864 208.906 URID052163 202.599 URID059279
200.919 URID046251 198.115 URID039539 195.616 URID016736
195.616 URID053062 195.616 URID062771 195.616 URID062818
193.941 URID049371 191.287 URID064743 187.454 URID055310
186.825 URID043732 185.338 URID019672 181.657 URID060155
181.657 URID061868 181.657 URID066010 181.657 URID070954
180.643 URID042029 179.218 URID039904 179.218 URID051924
179.218 URID055867 179.218 URID057439 177.330 URID049380
177.330 URID053404 177.330 URID059281 175.378 URID013603
175.313 URID046918 175.313 URID050365 175.313 URID053476
175.313 URID053769 175.313 URID058745 175.313 URID059432
175.313 URID066748 175.077 URID039332 172.607 URID032937
172.607 URID064447 171.970 URID013026 170.554 URID056265
170.554 URID061112 161.869 URID049596 161.478 URID017396
161.478 URID037132 161.478 URID050313 157.951 URID043698
```

154.032 URID046539 154.032 URID052654 154.032 URID062201
151.062 URID061742 149.015 URID034152 145.875 URID042869
143.636 URID056266 142.487 URID056148 142.487 URID060751
142.103 URID047754 142.103 URID050565 142.103 URID052738
142.103 URID061878 142.040 URID065568 139.692 URID045411
139.692 URID055881 139.692 URID056238 138.727 URID009007
137.065 URID048607 136.730 URID027392 136.730 URID032215
136.242 URID070979 133.982 URID051244 133.411 URID023902
129.894 URID045982 129.752 URID050981 129.752 URID055377
129.752 URID063378 129.752 URID066822 129.580 URID019456
129.580 URID035433 129.013 URID072104 128.652 URID044643
128.335 URID060474 128.110 URID049633 126.815 URID012268
126.730 URID050516 124.754 URID022408 124.754 URID043160
123.924 URID036451 123.684 URID005277 123.445 URID038320
123.021 URID038731 122.756 URID043087 121.593 URID016951
119.837 URID051566 119.478 URID048019 119.478 URID050725
119.478 URID052372 119.478 URID055096 119.478 URID064923
116.787 URID041231 116.718 URID021573 116.718 URID050312
116.718 URID053292 116.718 URID053464 116.718 URID066025
116.224 URID054566 115.761 URID040244 114.056 URID023544
113.258 URID009053 112.455 URID052115 111.939 URID058630
111.236 URID021278 110.284 URID054941 106.284 URID050466
105.974 URID053815 104.769 URID055012 103.962 URID045807
103.904 URID020850 102.744 URID025420 101.781 URID038472
101.264 URID035182 100.708 URID071145 99.2654 URID051286
97.8082 URID027109 97.8082 URID040236 97.8082 URID044644
97.8082 URID048051 97.8082 URID060509 97.8082 URID060772
97.8082 URID062769 97.8082 URID067241 97.4172 URID048821
96.3781 URID028309 96.2048 URID048371 96.2048 URID052646
95.9597 URID069509 95.6433 URID027195 95.6433 URID050295
95.3084 URID044611 94.6304 URID037082 94.6304 URID037624
94.6304 URID044132 94.6304 URID045862 94.6304 URID050728
94.6304 URID051572 94.6304 URID052948 94.6304 URID065989
93.8154 URID043446 91.8132 URID028519 91.3256 URID058069
91.3256 URID058245 90.8283 URID051886 90.8283 URID064791
90.2781 URID033961 88.6649 URID035604 88.6649 URID043295
88.6649 URID060621 88.6649 URID008964 87.6565 URID052195
87.6565 URID052262 87.6565 URID053410 87.6565 URID053774
87.6565 URID054323 87.6565 URID055769 87.6565 URID056691
87.6565 URID059144 87.6565 URID061188 87.6565 URID061749
87.6565 URID066756 87.6523 URID028045 86.6300 URID028782
86.3033 URID015516 86.3033 URID025096 86.3033 URID029259
86.3033 URID029387 86.3033 URID050029 86.3033 URID063495
86.1050 URID024488 86.1050 URID031978 86.1050 URID039588
86.0090 URID070656 86.0090 URID071290 86.0090 URID071649
86.0090 URID072682 85.2768 URID029596 85.2768 URID030240
85.2768 URID033813 85.2768 URID037127 85.2768 URID045434
85.2768 URID047064 85.2768 URID049408 85.2768 URID059254
85.2768 URID062525 85.2768 URID064305 85.2768 URID006544
85.2768 URID071118 84.3176 URID057058 83.2339 URID063955
83.1695 URID051727 83.1695 URID060503 83.1695 URID063136
83.1695 URID067335 81.8324 URID033168 80.2921 URID042959
79.8265 URID029161 78.1981 URID017959 77.7242 URID067238
77.0158 URID006306 77.0158 URID063809 76.5869 URID047671
76.5869 URID052959 76.5869 URID057432 76.2284 URID019848
76.2284 URID045798 76.2284 URID058967 74.4490 URID044663
74.2107 URID001451 74.2107 URID031799 74.2107 URID004916

```
73.0057 URID063344 72.9792 URID043120 72.1904 URID049808
71.2433 URID020963 71.2433 URID031796 71.2433 URID043271
71.2433 URID043671 71.2433 URID049560 71.2433 URID052509
71.2433 URID052673 71.2433 URID054552 71.2433 URID055116
71.2433 URID055380 71.2433 URID060252 71.2433 URID060389
71.2433 URID062369 71.2433 URID062872 71.2433 URID066418
71.0515 URID034228 71.0515 URID046806 71.0515 URID048789
71.0515 URID051470 71.0515 URID056752 71.0515 URID057393
71.0515 URID062786 71.0515 URID064091 71.0515 URID065354
71.0515 URID065357 70.5251 URID039846 70.2043 URID035476
70.1008 URID067339 69.8458 URID020580 69.8458 URID057373
68.6655 URID038947 67.6626 URID041430 66.8480 URID060450
66.8480 URID062443 65.5062 URID035412 65.5062 URID048525
64.8758 URID025517 64.8758 URID053096 64.8751 URID037604
64.8751 URID043208 64.8751 URID050019 64.7210 URID035710
64.6561 URID032512 64.6561 URID046756 64.5221 URID048686
64.1365 URID031262 64.1365 URID041369 64.1365 URID051246
64.1365 URID055503 64.1365 URID063828 64.1365 URID063829
64.1365 URID065202 63.9732 URID067923 63.7835 URID044652
63.5624 URID027898 62.8770 URID042294 62.4254 URID039296
62.4254 URID049709 62.3771 URID023405 62.3771 URID043571
62.3771 URID048269 59.9185 URID001081 59.9185 URID001329
59.9185 URID002330 59.9185 URID030920 59.9185 URID032834
59.9185 URID004325 59.9185 URID048697 59.9185 URID050612
59.9185 URID050613 59.9185 URID052277 59.9185 URID005663
59.9185 URID061611 59.9185 URID062749 59.7392 URID012909
59.7392 URID014591 59.7392 URID014982 59.7392 URID016034
59.7392 URID018718 59.7392 URID019404 59.7392 URID024375
59.7392 URID024700 59.7392 URID026756 59.7392 URID027307
59.7392 URID031003 59.7392 URID032988 59.7392 URID036609
59.7392 URID040045 59.7392 URID040231 59.7392 URID040786
59.7392 URID041183 59.7392 URID041630 59.7392 URID045881
59.7392 URID046522 59.7392 URID046673 59.7392 URID046725
59.7392 URID048380 59.7392 URID049033 59.7392 URID050018
59.7392 URID050747 59.7392 URID051245 59.7392 URID054003
59.7392 URID054405 59.7392 URID055700 59.7392 URID057186
59.7392 URID057457 59.7392 URID060662 59.7392 URID060678
59.7392 URID061435 59.7392 URID062080 59.7392 URID006670
59.7392 URID067500 59.7392 URID067752 58.3590 URID030543
58.3590 URID036199 58.3590 URID036619 58.3590 URID054246
58.3590 URID055686 58.3590 URID061323 58.3590 URID061448
58.3590 URID062714 58.3590 URID071089 58.0221 URID064635
57.8807 URID023950 57.2390 URID060696 57.2390 URID063801 ) ;
```

# Appendix 6: Recall bases for all four quantification functions

The numbers given here are for the implicit relevance assessments (See Section 6.1.3)

The 'documents' column indicates the number of documents in which components have been retrieved and assessed for each topic. The 'relevant' column indicates the total relevance score for the assessed documents for each topic. The total relevance score is integers for the strict functions because components are either assigned 1.0 or zero – the relevance scores for the generalized functions may contain decimals since values less than 1 are assigned to some components (See the quantification functions in Figure 6.8 and Figure 6.9). The total relevance score is the sum of the relevance scores for each topic.

Note that there were no relevant documents for topic 50 in any of the strict functions.

| **Quantification function: gen_whole** | | | **Quantification function: str_whole** | | |
|---|---|---|---|---|---|
| Topic 31: | 49.51 relevant; | 988 documents | Topic 31: | 34.00 relevant; | 988 documents |
| Topic 32: | 329.86 relevant; | 932 documents | Topic 32: | 52.00 relevant; | 932 documents |
| Topic 33: | 42.79 relevant; | 986 documents | Topic 33: | 22.00 relevant; | 986 documents |
| Topic 34: | 364.47 relevant; | 1323 documents | Topic 34: | 81.00 relevant; | 1323 documents |
| Topic 36: | 133.69 relevant; | 1020 documents | Topic 36: | 70.00 relevant; | 1020 documents |
| Topic 37: | 583.59 relevant; | 762 documents | Topic 37: | 279.00 relevant; | 762 documents |
| Topic 38: | 869.10 relevant; | 1109 documents | Topic 38: | 351.00 relevant; | 1109 documents |
| Topic 39: | 239.08 relevant; | 880 documents | Topic 39: | 82.00 relevant; | 880 documents |
| Topic 40: | 241.52 relevant; | 981 documents | Topic 40: | 161.00 relevant; | 981 documents |
| Topic 41: | 181.21 relevant; | 927 documents | Topic 41: | 169.00 relevant; | 927 documents |
| Topic 42: | 334.40 relevant; | 1012 documents | Topic 42: | 209.00 relevant; | 1012 documents |
| Topic 43: | 50.80 relevant; | 1038 documents | Topic 43: | 31.00 relevant; | 1038 documents |
| Topic 44: | 95.63 relevant; | 1443 documents | Topic 44: | 59.00 relevant; | 1443 documents |
| Topic 45: | 507.07 relevant; | 983 documents | Topic 45: | 85.00 relevant; | 983 documents |
| Topic 46: | 241.25 relevant; | 749 documents | Topic 46: | 68.00 relevant; | 749 documents |
| Topic 47: | 104.83 relevant; | 1305 documents | Topic 47: | 22.00 relevant; | 1305 documents |
| Topic 48: | 198.20 relevant; | 933 documents | Topic 48: | 86.00 relevant; | 933 documents |
| Topic 49: | 134.80 relevant; | 819 documents | Topic 49: | 16.00 relevant; | 819 documents |
| Topic 50: | 424.71 relevant; | 902 documents | Topic 50: | 0.00 relevant; | 902 documents |
| Topic 51: | 215.95 relevant; | 978 documents | Topic 51: | 79.00 relevant; | 978 documents |
| Topic 52: | 74.17 relevant; | 1084 documents | Topic 52: | 25.00 relevant; | 1084 documents |
| Topic 53: | 376.42 relevant; | 915 documents | Topic 53: | 55.00 relevant; | 915 documents |
| Topic 58: | 818.58 relevant; | 1286 documents | Topic 58: | 513.00 relevant; | 1286 documents |
| Topic 60: | 717.90 relevant; | 1257 documents | Topic 60: | 477.00 relevant; | 1257 documents |
| Total relevance score: 7329.53 | | | Total relevance score: 3026.00 | | |

| **Quantification function: gen_inex** | | | **Quantification function: str_inex** | | |
|---|---|---|---|---|---|
| Topic 31: | 45.25 relevant; | 988 documents | Topic 31: | 4.00 relevant; | 988 documents |
| Topic 32: | 795.50 relevant; | 932 documents | Topic 32: | 35.00 relevant; | 932 documents |
| Topic 33: | 34.50 relevant; | 986 documents | Topic 33: | 2.00 relevant; | 986 documents |
| Topic 34: | 412.50 relevant; | 1323 documents | Topic 34: | 66.00 relevant; | 1323 documents |
| Topic 36: | 138.75 relevant; | 1020 documents | Topic 36: | 31.00 relevant; | 1020 documents |
| Topic 37: | 860.50 relevant; | 762 documents | Topic 37: | 138.00 relevant; | 762 documents |
| Topic 38: | 1304.00 relevant; | 1109 documents | Topic 38: | 111.00 relevant; | 1109 documents |
| Topic 39: | 277.25 relevant; | 880 documents | Topic 39: | 48.00 relevant; | 880 documents |
| Topic 40: | 232.50 relevant; | 981 documents | Topic 40: | 124.00 relevant; | 981 documents |
| Topic 41: | 159.00 relevant; | 927 documents | Topic 41: | 57.00 relevant; | 927 documents |
| Topic 42: | 309.50 relevant; | 1012 documents | Topic 42: | 91.00 relevant; | 1012 documents |
| Topic 43: | 77.75 relevant; | 1038 documents | Topic 43: | 15.00 relevant; | 1038 documents |
| Topic 44: | 158.00 relevant; | 1443 documents | Topic 44: | 36.00 relevant; | 1443 documents |
| Topic 45: | 535.75 relevant; | 983 documents | Topic 45: | 57.00 relevant; | 983 documents |
| Topic 46: | 239.50 relevant; | 749 documents | Topic 46: | 26.00 relevant; | 749 documents |
| Topic 47: | 233.75 relevant; | 1305 documents | Topic 47: | 22.00 relevant; | 1305 documents |
| Topic 48: | 296.75 relevant; | 933 documents | Topic 48: | 65.00 relevant; | 933 documents |
| Topic 49: | 157.25 relevant; | 819 documents | Topic 49: | 9.00 relevant; | 819 documents |
| Topic 50: | 451.50 relevant; | 902 documents | Topic 50: | 0.00 relevant; | 902 documents |
| Topic 51: | 191.25 relevant; | 978 documents | Topic 51: | 26.00 relevant; | 978 documents |
| Topic 52: | 140.50 relevant; | 1084 documents | Topic 52: | 15.00 relevant; | 1084 documents |
| Topic 53: | 816.25 relevant; | 915 documents | Topic 53: | 34.00 relevant; | 915 documents |
| Topic 58: | 722.75 relevant; | 1286 documents | Topic 58: | 210.00 relevant; | 1286 documents |
| Topic 60: | 638.50 relevant; | 1257 documents | Topic 60: | 174.00 relevant; | 1257 documents |
| Total relevance score: 9228.75 | | | Total relevance score: 1396.00 | | |

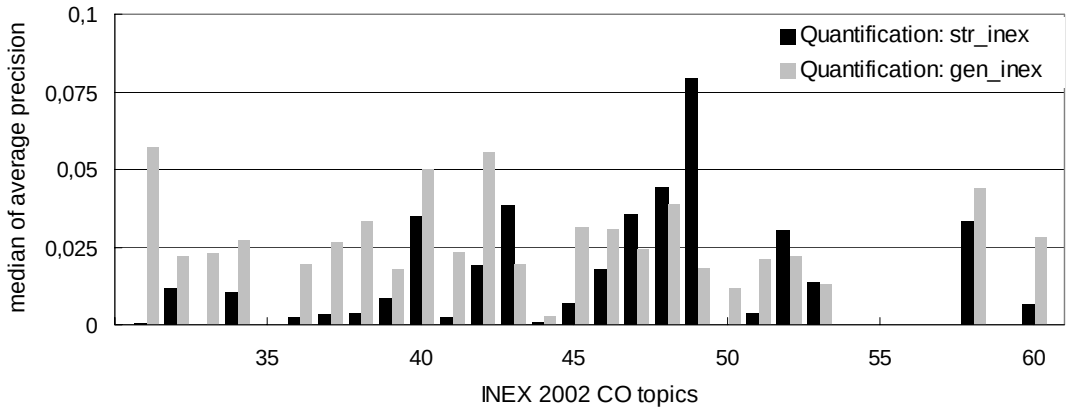
# Appendix 7: Details of the official INEX 2002 runs.

The next three pages reproduce the official results from the INEX 2002 workshop proceedings (Fuhr et al., 2003) of the three CO runs that were submitted to INEX 2002 as part of the TAPIR project:

- A "bag-of-words" run, corresponding to baseline 2.
- A "boomerang" run, corresponding to a best match boomerang effect run with DCV_step1 = 500, extended citation indexes in both Step 2 and Step 3 (xx), and with no CCV_step2 threshold invoked.
- A "polyrepresentation" run with DCV_step1 = 500.

The following details are presented for each run. Results are given for str_inex and gen_inex (simply called "strict" and "generalized"):

- A P-R curve, plotting the precision values over 100 recall points, including a smaller diagram plotting the run together with the other 48 submissions for comparison.
- The average precision value calculated over those 100 recall points.
- A table displaying the average precision values for each topic.
- A topic-by-topic histogram that compares the average precision of each topic against the median performance of all runs in INEX 2002 (The figure below shows an overview of the median average precision values per topic).



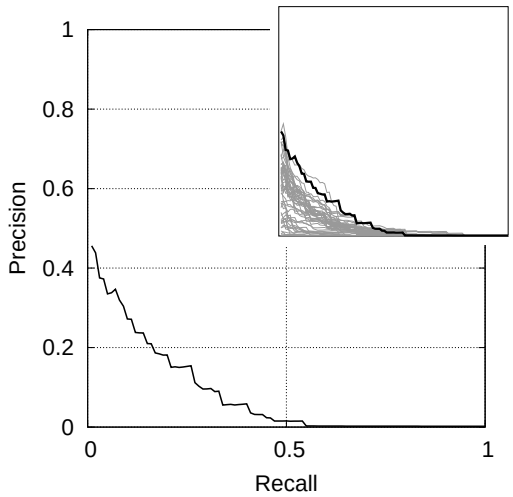

**Overview of the median of average precision for all 49 runs submitted INEX 2002 CO. (Source: Redrawing of Figure b in Fuhr et al. (2003, p. 188). The data for this was kindly provided by Norbert Gövert of University of Duisberg, Germany.)**

The results presented here may deviate slightly from the ones presented in Chapter 6. This is due to the correction of 60 erroneous document IDs that were corrected after the submission to INEX 2002.

# Royal School of Library and Information Science
# bag-of-words (CO)

## Quantisation: strict

**Recall / precision graph:**

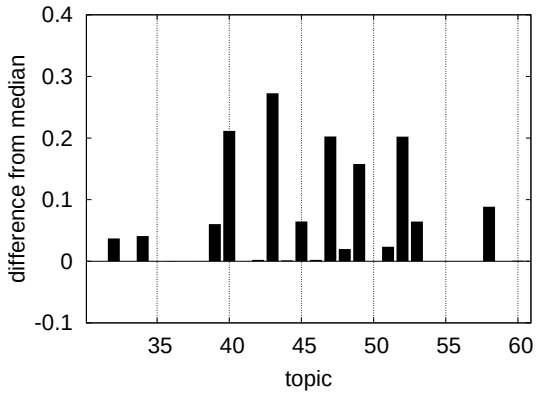



**Overall average precision:** 0.0809

**Average precision per topic:**

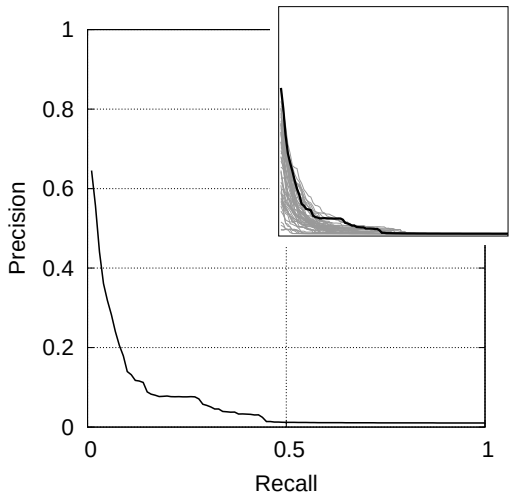
| | | | | | |
|---|---|---|---|---|---|
| 31 | 0.0002 | 41 | 0.0024 | 51 | 0.0275 |
| 32 | 0.0487 | 42 | 0.0215 | 52 | 0.2325 |
| 33 | 0.0001 | 43 | 0.3109 | 53 | 0.0780 |
| 34 | 0.0511 | 44 | 0.0022 | 54 | – |
| 35 | – | 45 | 0.0715 | 55 | – |
| 36 | 0.0021 | 46 | 0.0201 | 56 | – |
| 37 | 0.0032 | 47 | 0.2379 | 57 | – |
| 38 | 0.0039 | 48 | 0.0641 | 58 | 0.1219 |
| 39 | 0.0689 | 49 | 0.2376 | 59 | – |
| 40 | 0.2465 | 50 | – | 60 | 0.0077 |

**Difference from median in average precision per topic:**




## Quantisation: generalised

**Recall / precision graph:**

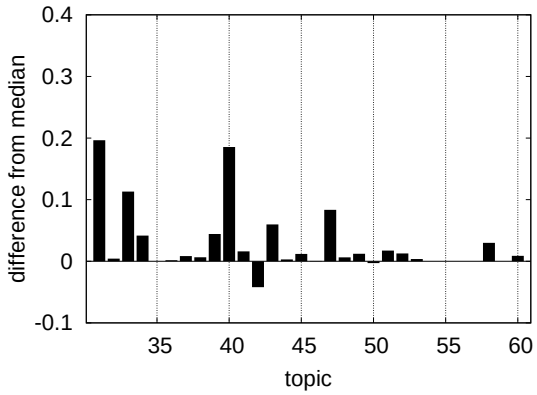



**Overall average precision:** 0.0618

**Average precision per topic:**

| | | | | | |
|---|---|---|---|---|---|
| 31 | 0.2535 | 41 | 0.0392 | 51 | 0.0386 |
| 32 | 0.0262 | 42 | 0.0134 | 52 | 0.0349 |
| 33 | 0.1360 | 43 | 0.0793 | 53 | 0.0168 |
| 34 | 0.0688 | 44 | 0.0059 | 54 | – |
| 35 | – | 45 | 0.0432 | 55 | – |
| 36 | 0.0213 | 46 | 0.0312 | 56 | – |
| 37 | 0.0351 | 47 | 0.1079 | 57 | – |
| 38 | 0.0395 | 48 | 0.0449 | 58 | 0.0739 |
| 39 | 0.0620 | 49 | 0.0305 | 59 | – |
| 40 | 0.2356 | 50 | 0.0085 | 60 | 0.0367 |

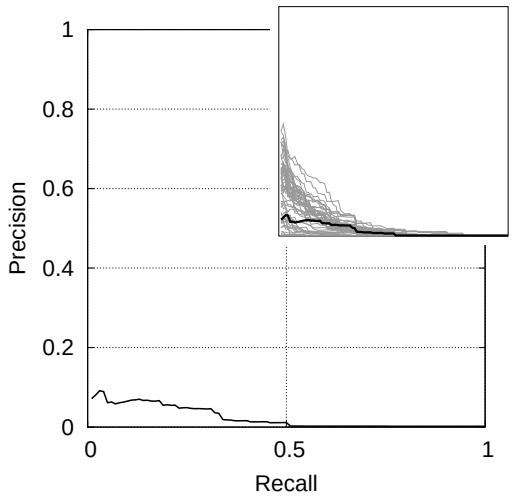
**Difference from median in average precision per topic:**



# Royal School of Library and Information Science
# boomerang (CO)

## Quantisation: strict

**Recall / precision graph:**

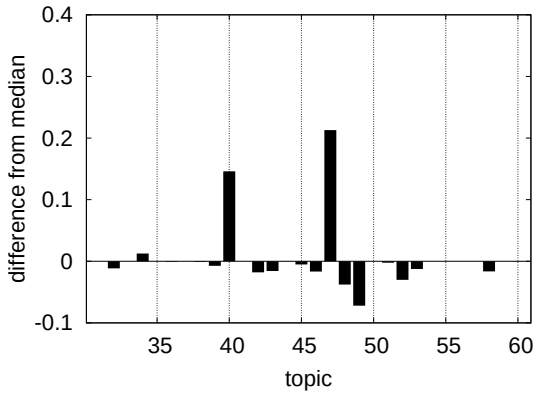



**Overall average precision:** 0.0231

**Average precision per topic:**

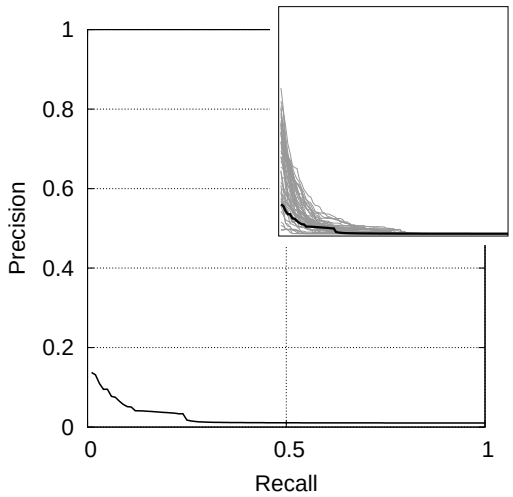
| | | | | | |
|---|---|---|---|---|---|
| 31 | 0.0002 | 41 | 0.0024 | 51 | 0.0016 |
| 32 | 0.0002 | 42 | 0.0014 | 52 | 0.0001 |
| 33 | 0.0001 | 43 | 0.0225 | 53 | 0.0011 |
| 34 | 0.0228 | 44 | 0.0005 | 54 | – |
| 35 | – | 45 | 0.0017 | 55 | – |
| 36 | 0.0017 | 46 | 0.0012 | 56 | – |
| 37 | 0.0032 | 47 | 0.2482 | 57 | – |
| 38 | 0.0026 | 48 | 0.0065 | 58 | 0.0170 |
| 39 | 0.0012 | 49 | 0.0076 | 59 | – |
| 40 | 0.1810 | 50 | – | 60 | 0.0065 |

**Difference from median in average precision per topic:**




## Quantisation: generalised

**Recall / precision graph:**

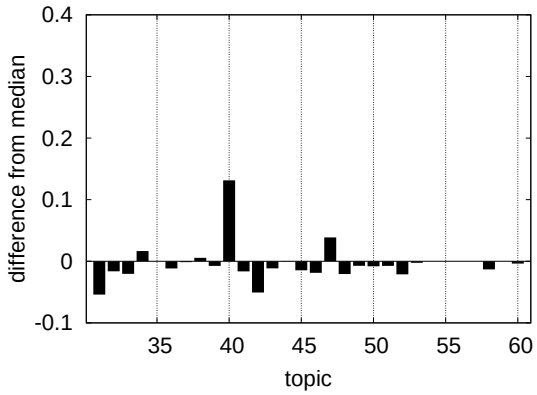



**Overall average precision:** 0.0227

**Average precision per topic:**

| | | | | | |
|---|---|---|---|---|---|
| 31 | 0.0029 | 41 | 0.0068 | 51 | 0.0137 |
| 32 | 0.0056 | 42 | 0.0048 | 52 | 0.0010 |
| 33 | 0.0025 | 43 | 0.0080 | 53 | 0.0107 |
| 34 | 0.0437 | 44 | 0.0024 | 54 | – |
| 35 | – | 45 | 0.0166 | 55 | – |
| 36 | 0.0080 | 46 | 0.0120 | 56 | – |
| 37 | 0.0253 | 47 | 0.0631 | 57 | – |
| 38 | 0.0387 | 48 | 0.0179 | 58 | 0.0309 |
| 39 | 0.0103 | 49 | 0.0111 | 59 | – |
| 40 | 0.1815 | 50 | 0.0034 | 60 | 0.0241 |

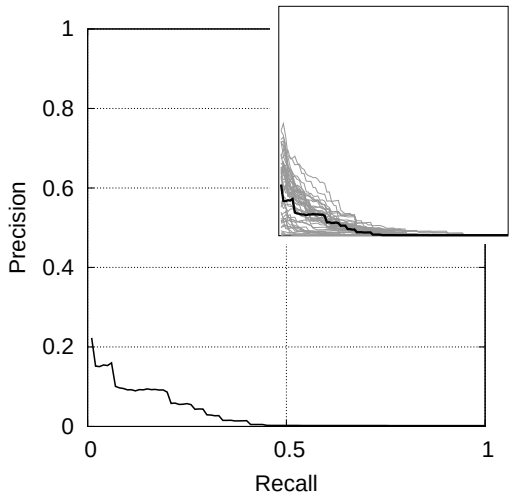
**Difference from median in average precision per topic:**



# Royal School of Library and Information Science
# polyrepresentation (CO)

## Quantisation: strict

**Recall / precision graph:**

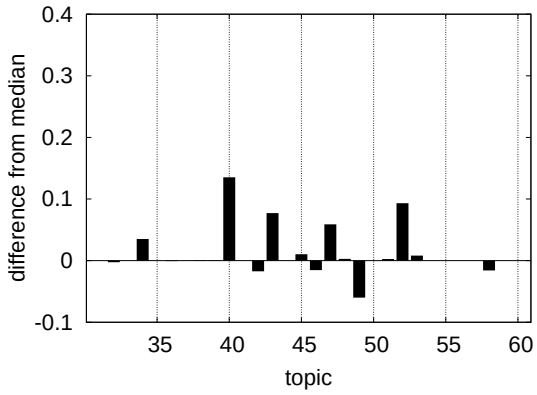



**Overall average precision:** 0.0313

**Average precision per topic:**

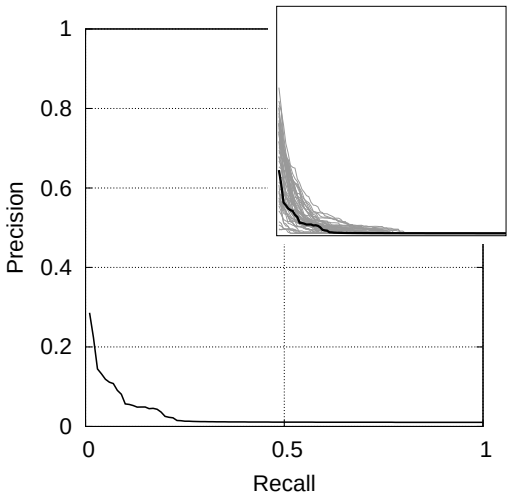
| | | | | | |
|---|---|---|---|---|---|
| 31 | 0.0002 | 41 | 0.0024 | 51 | 0.0065 |
| 32 | 0.0091 | 42 | 0.0021 | 52 | 0.1234 |
| 33 | 0.0001 | 43 | 0.1154 | 53 | 0.0218 |
| 34 | 0.0453 | 44 | 0.0005 | 54 | – |
| 35 | – | 45 | 0.0173 | 55 | – |
| 36 | 0.0017 | 46 | 0.0026 | 56 | – |
| 37 | 0.0032 | 47 | 0.0943 | 57 | – |
| 38 | 0.0044 | 48 | 0.0472 | 58 | 0.0174 |
| 39 | 0.0080 | 49 | 0.0197 | 59 | – |
| 40 | 0.1702 | 50 | – | 60 | 0.0066 |

**Difference from median in average precision per topic:**




## Quantisation: generalised

**Recall / precision graph:**

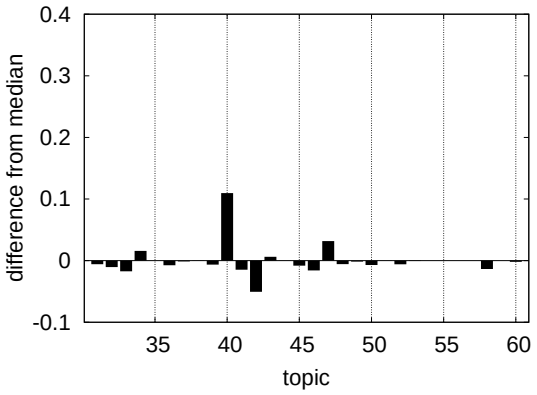



**Overall average precision:** 0.0271

**Average precision per topic:**

| | | | | | |
|---|---|---|---|---|---|
| 31 | 0.0512 | 41 | 0.0085 | 51 | 0.0210 |
| 32 | 0.0115 | 42 | 0.0050 | 52 | 0.0163 |
| 33 | 0.0055 | 43 | 0.0257 | 53 | 0.0126 |
| 34 | 0.0431 | 44 | 0.0031 | 54 | – |
| 35 | – | 45 | 0.0231 | 55 | – |
| 36 | 0.0121 | 46 | 0.0148 | 56 | – |
| 37 | 0.0253 | 47 | 0.0561 | 57 | – |
| 38 | 0.0327 | 48 | 0.0329 | 58 | 0.0305 |
| 39 | 0.0113 | 49 | 0.0168 | 59 | – |
| 40 | 0.1597 | 50 | 0.0044 | 60 | 0.0260 |

**Difference from median in average precision per topic:**



# Appendix 8: Enlarged versions of the P-R curves.

This appendix contains enlarged versions of the following curves:

Figure 7.1.a-d. P-R curves of the three official runs submitted to INEX 2002 for the four quantification functions.

Figure 7.2.a-d. P-R curves of the best match boomerang effect and the baseline runs as tested in research question 1 for the four quantification functions. Note that the y-axis is fitted to each curve.

Figure 7.7.a-d. P-R curves of all DCV_step1 values of a particular combination of CCV_step2 (High) and citation indexes (xf) for the gen_whole quantification function. The best absolute AvgP run (DCV_step1 = 16) is shown together with the runs that were not significantly different from it (a and b), as well the runs that were (c and d). Note that only up to 0.50 recall is shown.

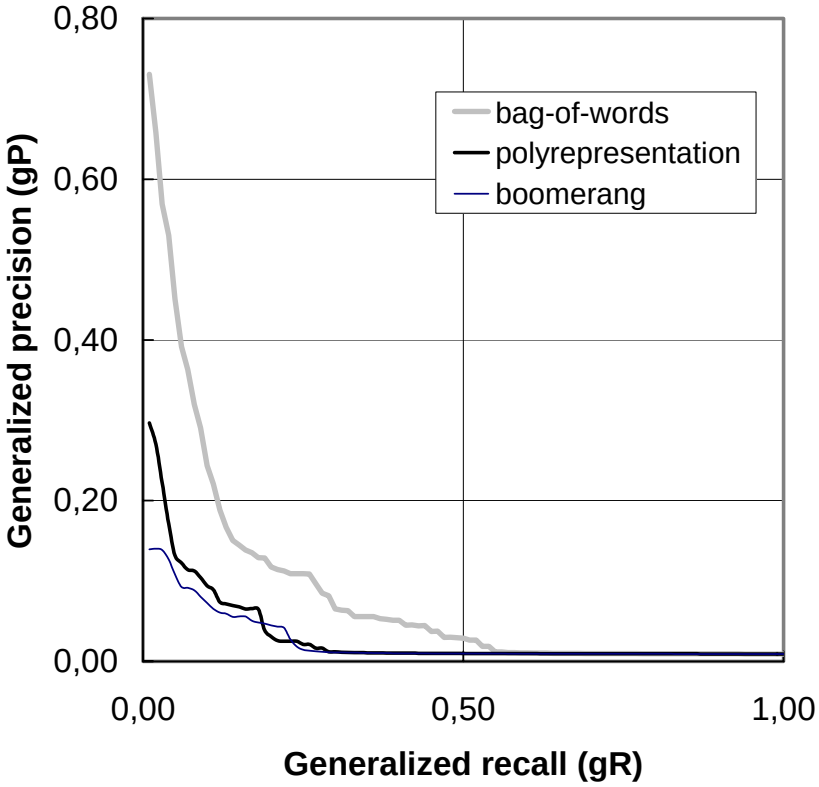



Figure 7.1.a (p. 172):

P-R curves for the three official runs submitted to INEX 2002 for the four quantification functions. Note that the y-axis is fitted to each curve.

Quantification function: gen_whole

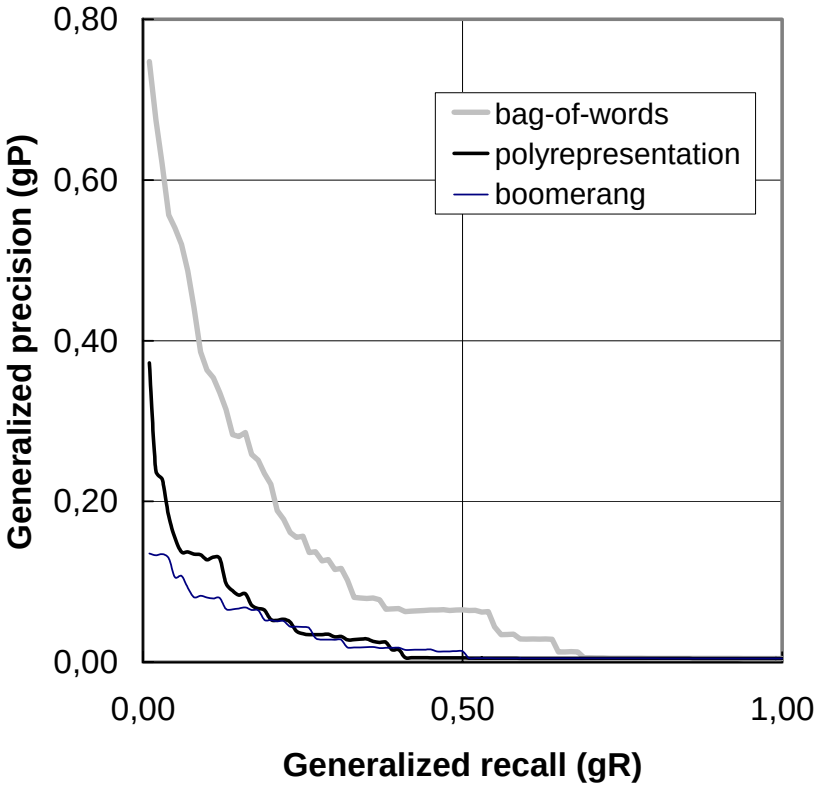



Figure 7.1.b (p. 172):

P-R curves for the three official runs submitted to INEX 2002 for the four quantification functions. Note that the y-axis is fitted to each curve.

Quantification function: str_whole

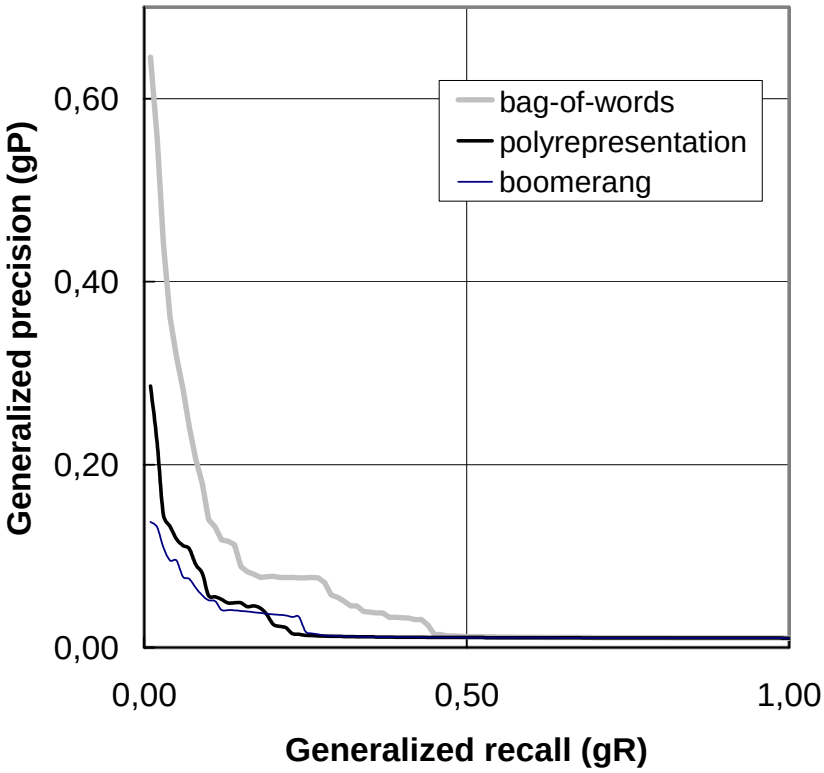



Figure 7.1.c (p. 172):

P-R curves for the three official runs submitted to INEX 2002 for the four quantification functions. Note that the y-axis is fitted to each curve.

Quantification function: gen_inex

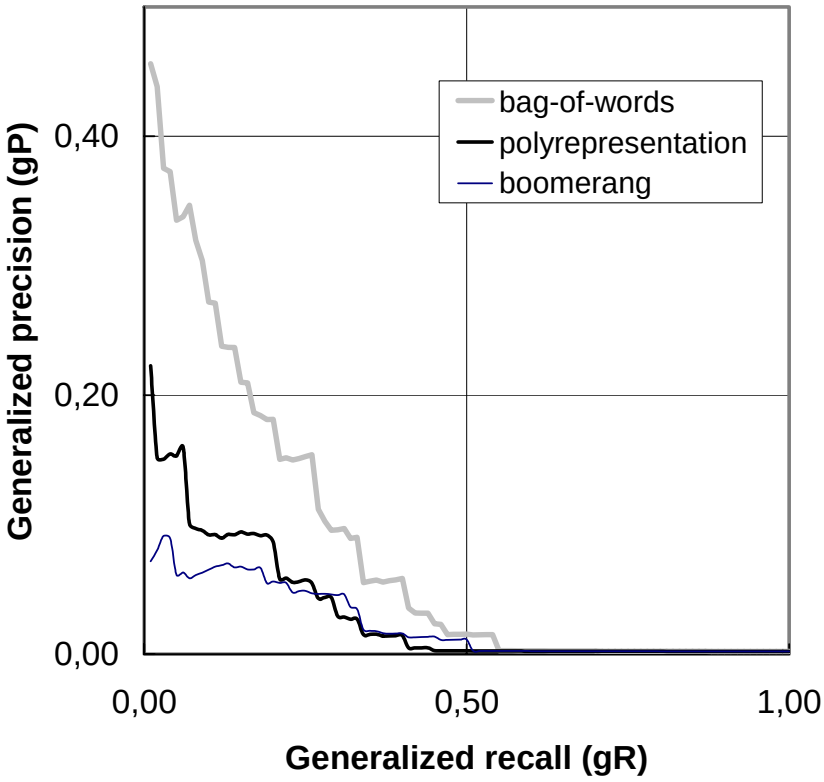



Figure 7.1.d (p. 172):

P-R curves for the three official runs submitted to INEX 2002 for the four quantification functions. Note that the y-axis is fitted to each curve.

Quantification function: str_inex

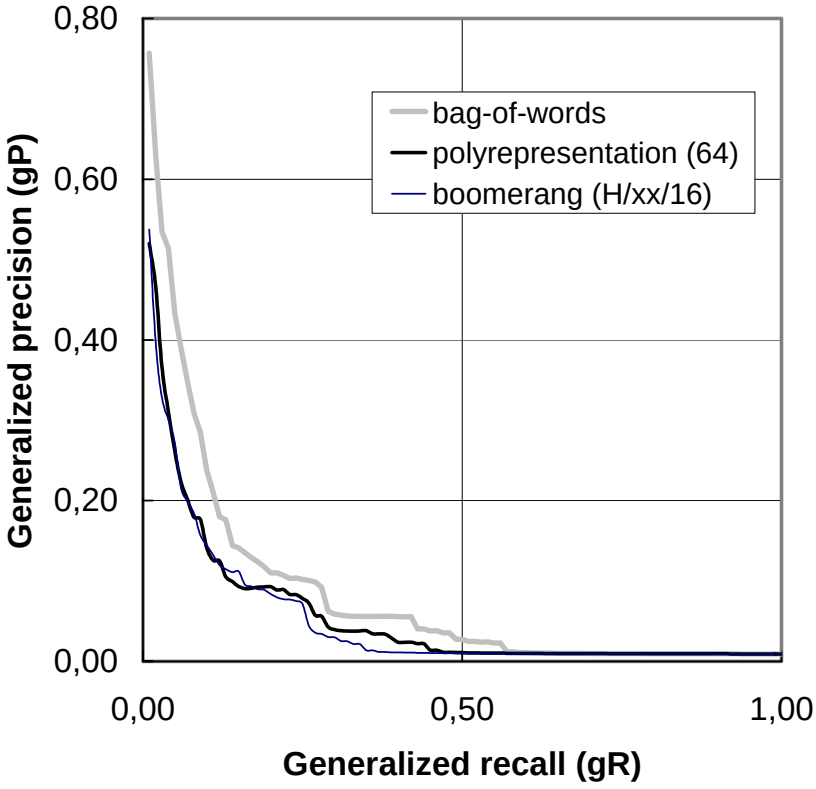



Figure 7.2.a (p. 189):

P-R curves of the three runs tested in research question 1 for the four quantification functions.

Quantification function: gen_whole

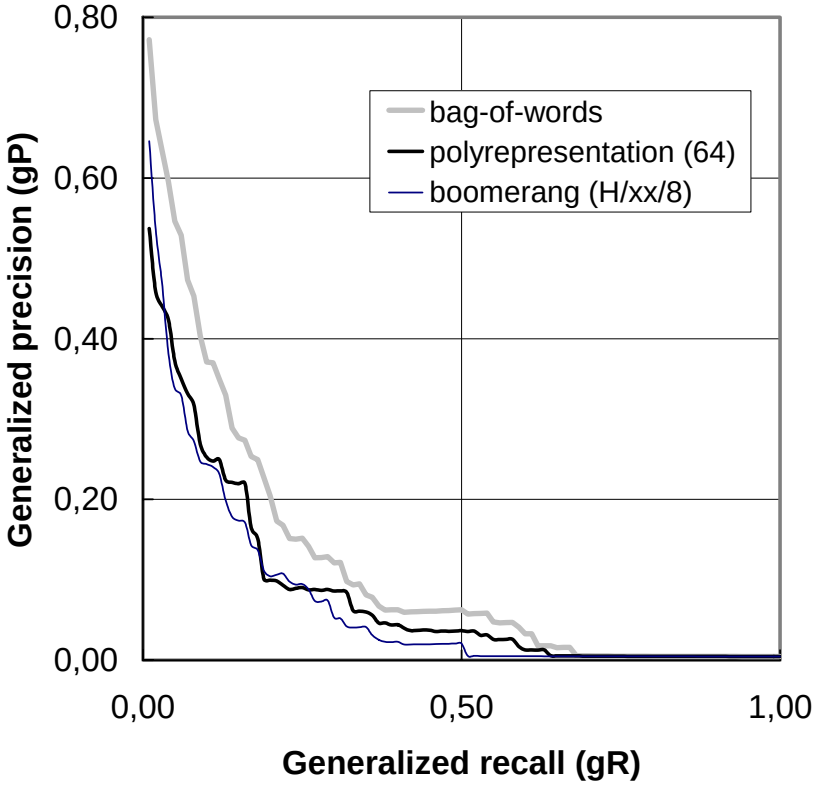



Figure 7.2.b (p. 189):

P-R curves of the three runs tested in research question 1 for the four quantification functions.

Quantification function: str_whole

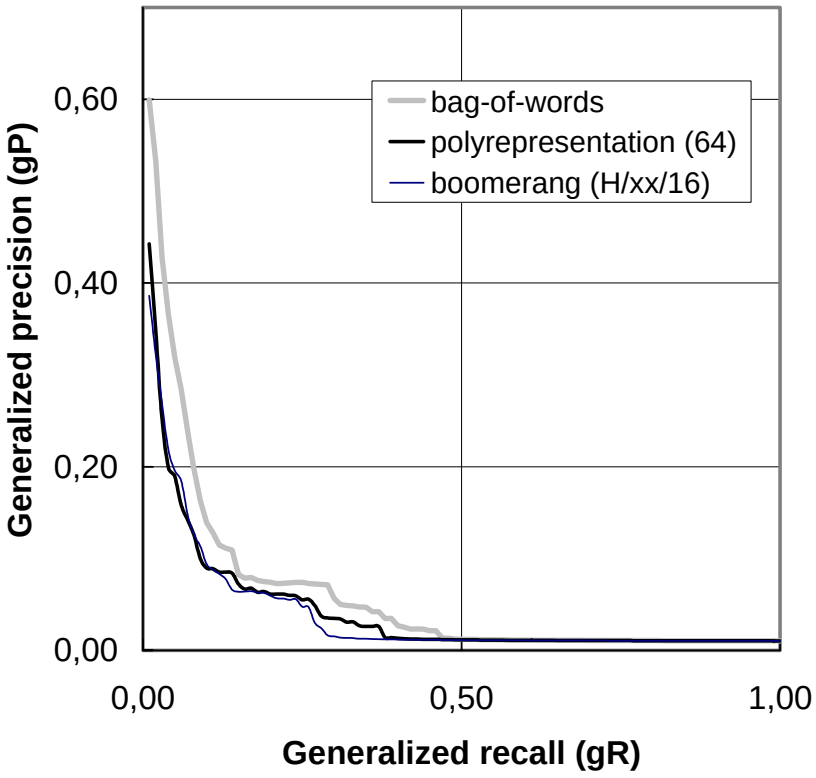



Figure 7.2.c (p. 189):

P-R curves of the three runs tested in research question 1 for the four quantification functions.

Quantification function: gen_inex

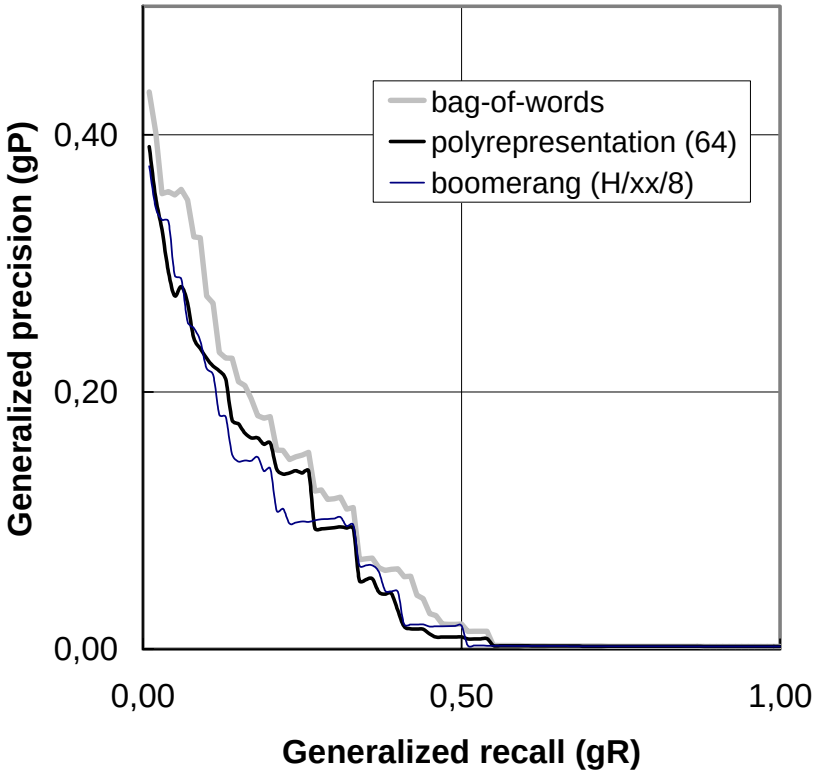



Figure 7.2.d (p. 189):

P-R curves of the three runs tested in research question 1 for the four quantification functions.

Quantification function: str_inex

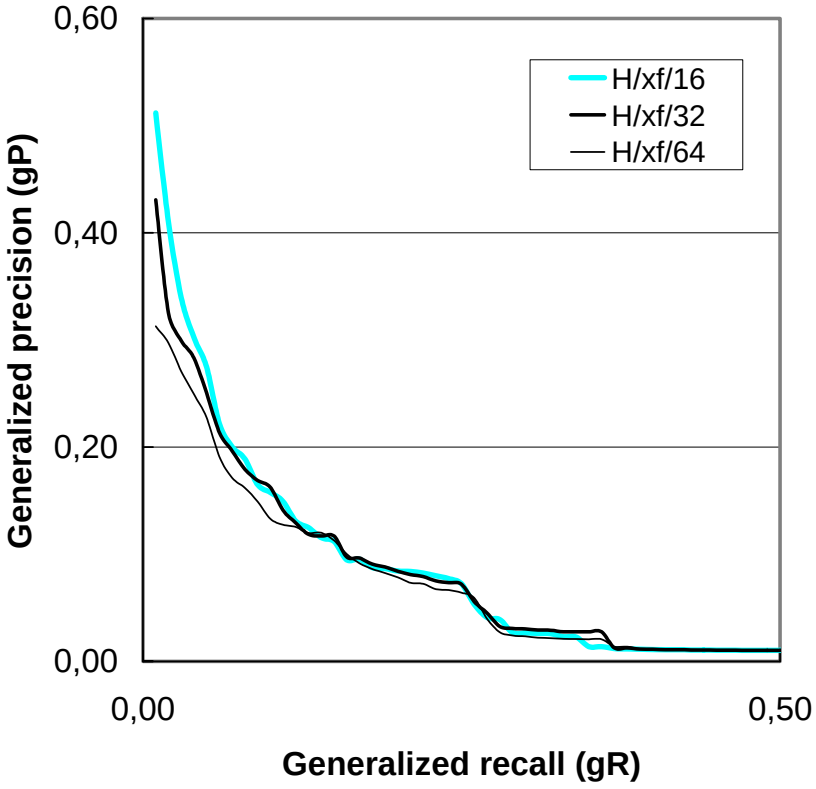



Figure 7.7.a (p.198):

P-R curves of all DCV_step1 values of a particular combination of CCV_step2 (High) and citation indexes (xf) for the gen_whole quantification function. The best absolute AvgP run (DCV_step1 = 16) is shown together with the runs that were not significantly different from it (a and b), as well the runs that were (c and d). Note that only up to 0.50 recall is shown.

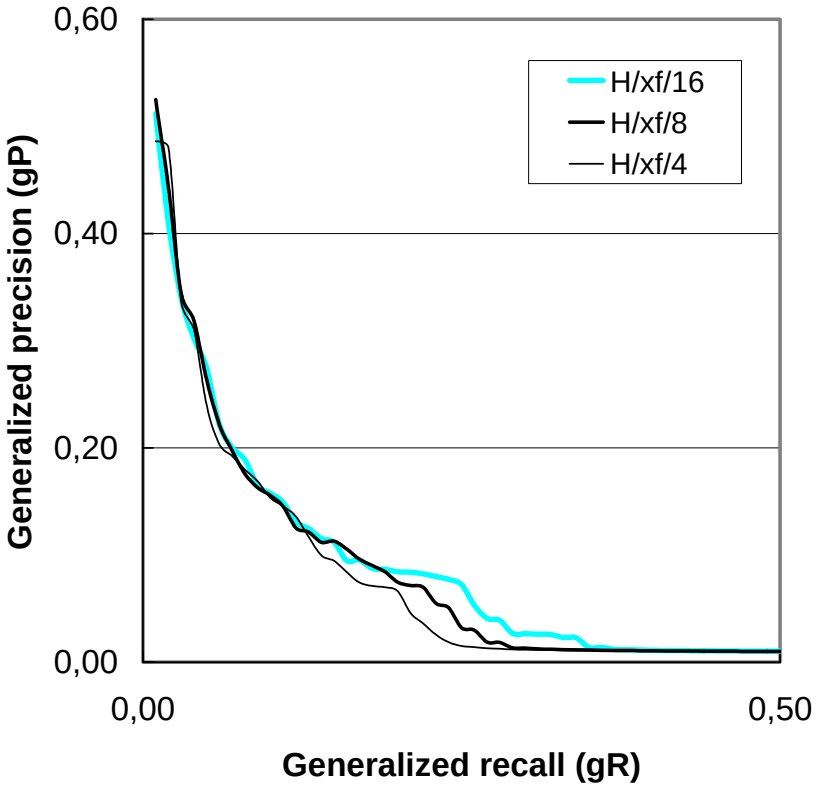



Figure 7.7.b (p.198):

P-R curves of all DCV_step1 values of a particular combination of CCV_step2 (High) and citation indexes (xf) for the gen_whole quantification function. The best absolute AvgP run (DCV_step1 = 16) is shown together with the runs that were not significantly different from it (a and b), as well the runs that were (c and d). Note that only up to 0.50 recall is shown.

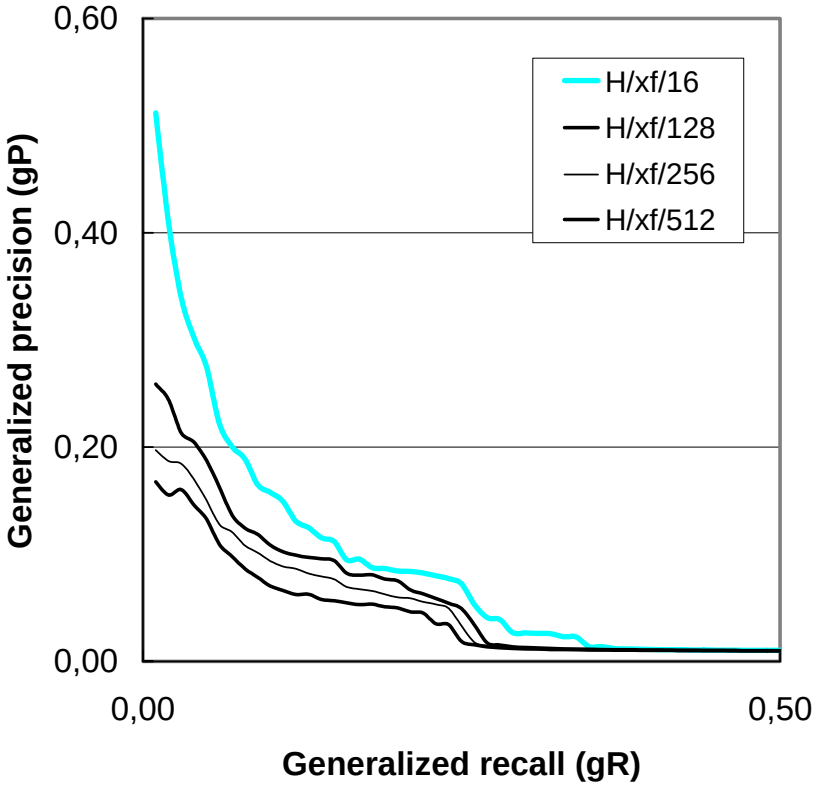



Figure 7.7.c (p.198):

P-R curves of all DCV_step1 values of a particular combination of CCV_step2 (High) and citation indexes (xf) for the gen_whole quantification function. The best absolute AvgP run (DCV_step1 = 16) is shown together with the runs that were not significantly different from it (a and b), as well the runs that were (c and d). Note that only up to 0.50 recall is shown.

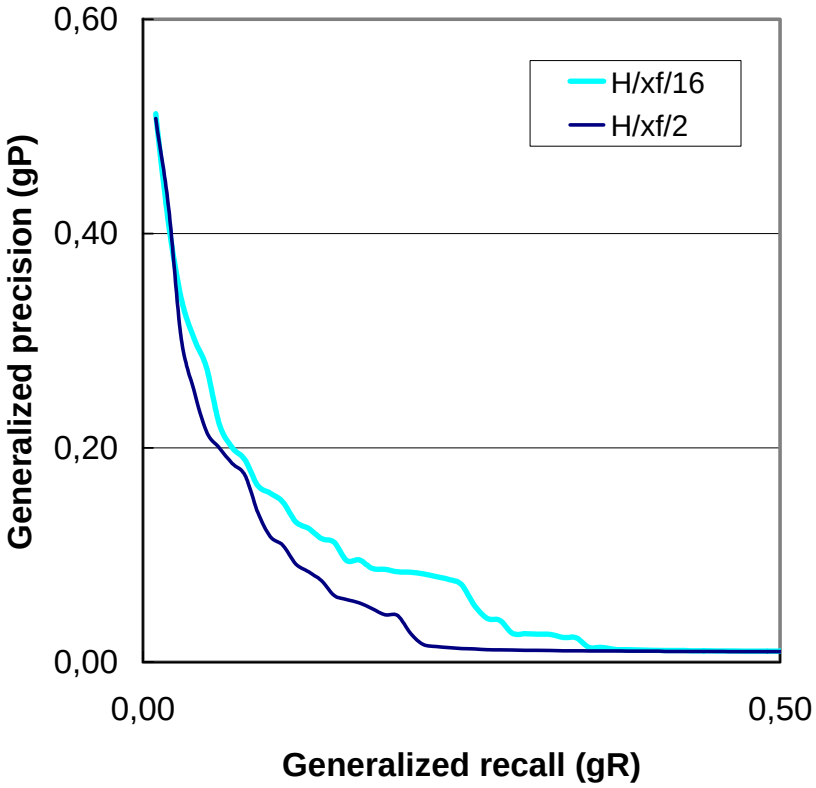



Figure 7.7.d (p.198):

P-R curves of all DCV_step1 values of a particular combination of CCV_step2 (High) and citation indexes (xf) for the gen_whole quantification function. The best absolute AvgP run (DCV_step1 = 16) is shown together with the runs that were not significantly different from it (a and b), as well the runs that were (c and d). Note that only up to 0.50 recall is shown.

# Appendix 9: Enlarged versions of the (nD)CG curves.

This appendix contains enlarged versions of the following curves:

Figure 7.3.a-d. (nD)CG curves of the best match boomerang effect and the baseline runs as investigated in research question 1. Gain values for the computations were 0-1-2-3, and the natural logarithm (2.718) was used as discount factor for the (n)DCG computations.

Figure 7.4.a-d. (nD)CG curves of the best match boomerang effect and the baseline runs as investigated in research question 1. Gain values for the computations were 0-0-0-3, and the natural logarithm (2.718) was used as discount factor for the (n)DCG computations.

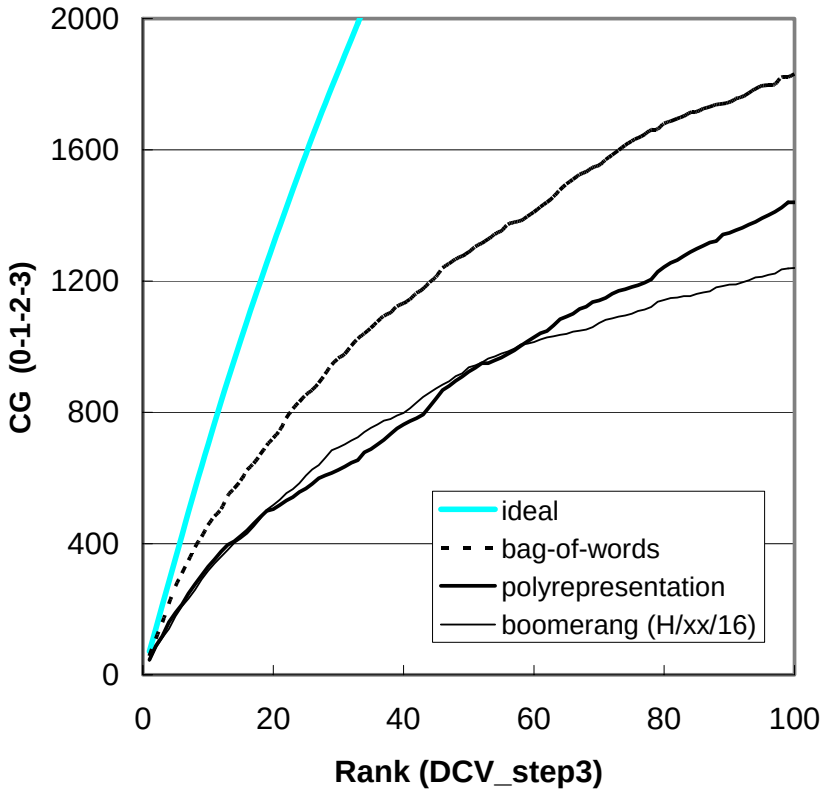



Figure 7.3.a (p. 190):

(nD)CG curves of the best match boomerang effect and the baseline runs as investigated in research question 1. Gain values for the computations were 0-1-2-3, and the natural logarithm was used as discount factor for the (n)DCG computations.

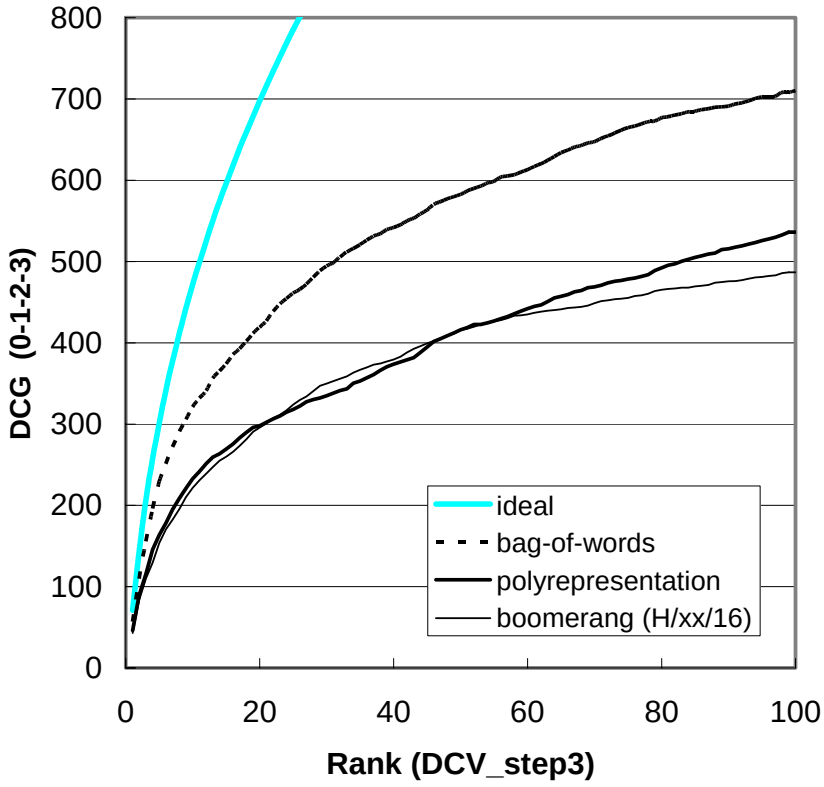



Figure 7.3.b (p. 190):

(nD)CG curves of the best match boomerang effect and the baseline runs as investigated in research question 1. Gain values for the computations were 0-1-2-3, and the natural logarithm was used as discount factor for the (n)DCG computations.

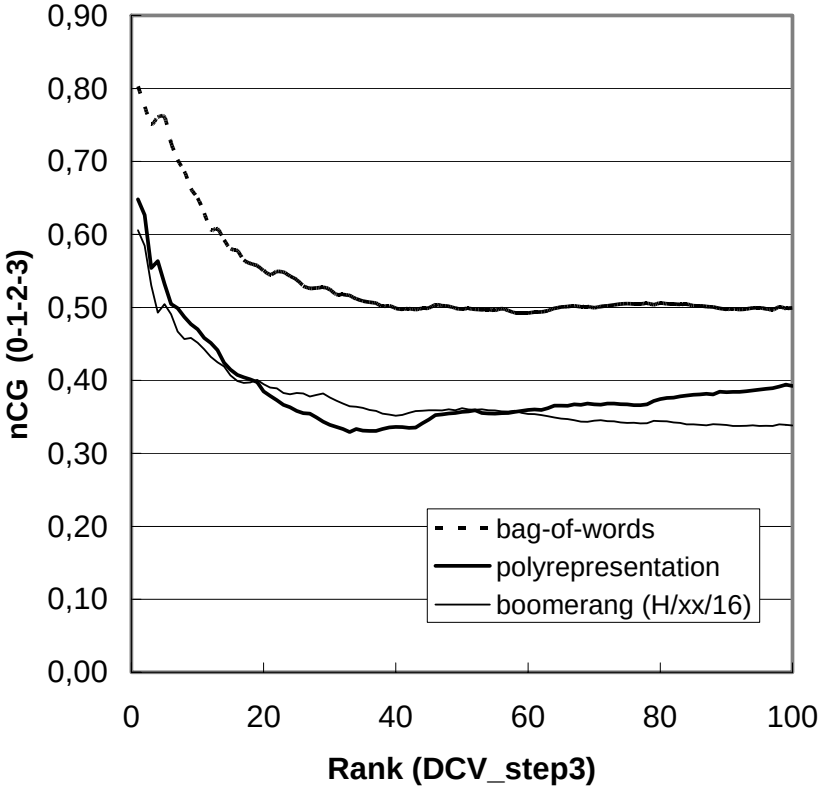



Figure 7.3.c (p. 190):

(nD)CG curves of the best match boomerang effect and the baseline runs as investigated in research question 1. Gain values for the computations were 0-1-2-3, and the natural logarithm was used as discount factor for the (n)DCG computations.

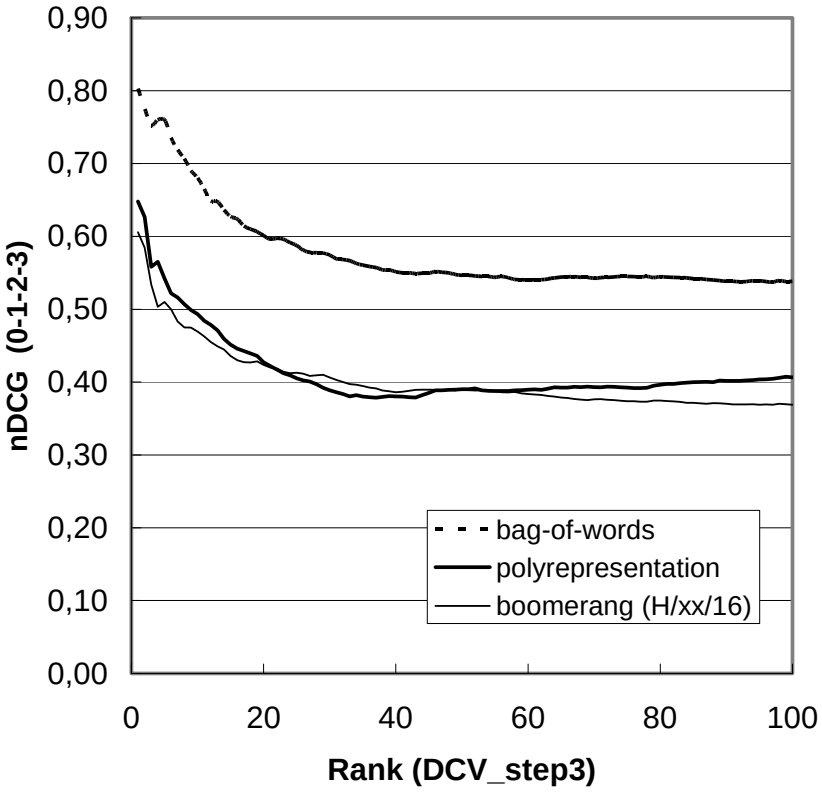



Figure 7.3.d (p. 190):

(nD)CG curves of the best match boomerang effect and the baseline runs as investigated in research question 1. Gain values for the computations were 0-1-2-3, and the natural logarithm was used as discount factor for the (n)DCG computations.

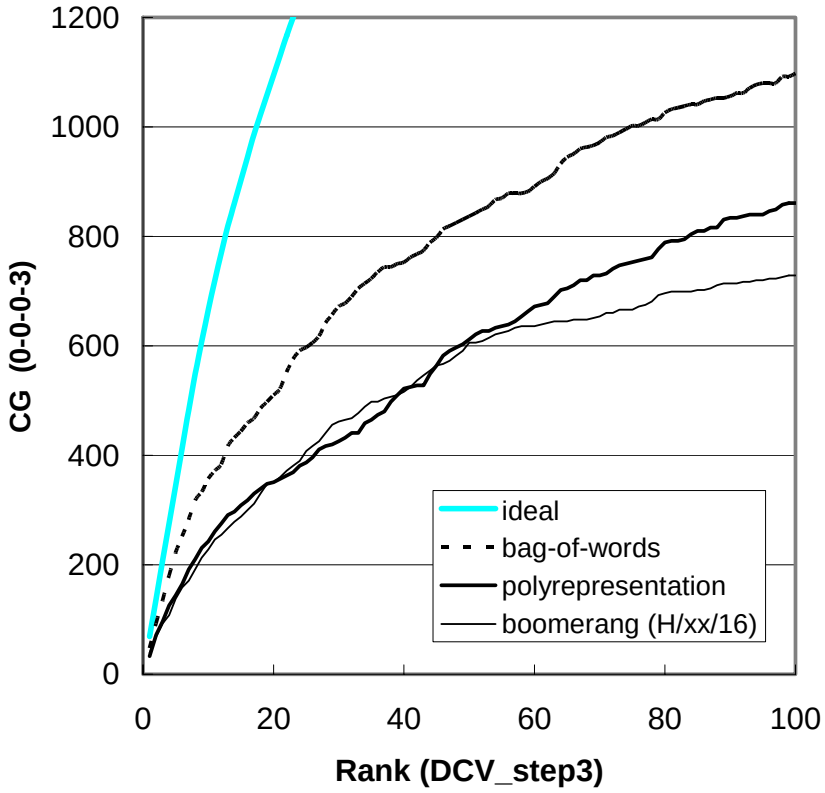



Figure 7.4.a (p. 191):

(nD)CG curves of the best match boomerang effect and the baseline runs as investigated in research question 1. Gain values for the computations were 0-0-0-3, and the natural logarithm (2.718) was used as discount factor for the (n)DCG computations.

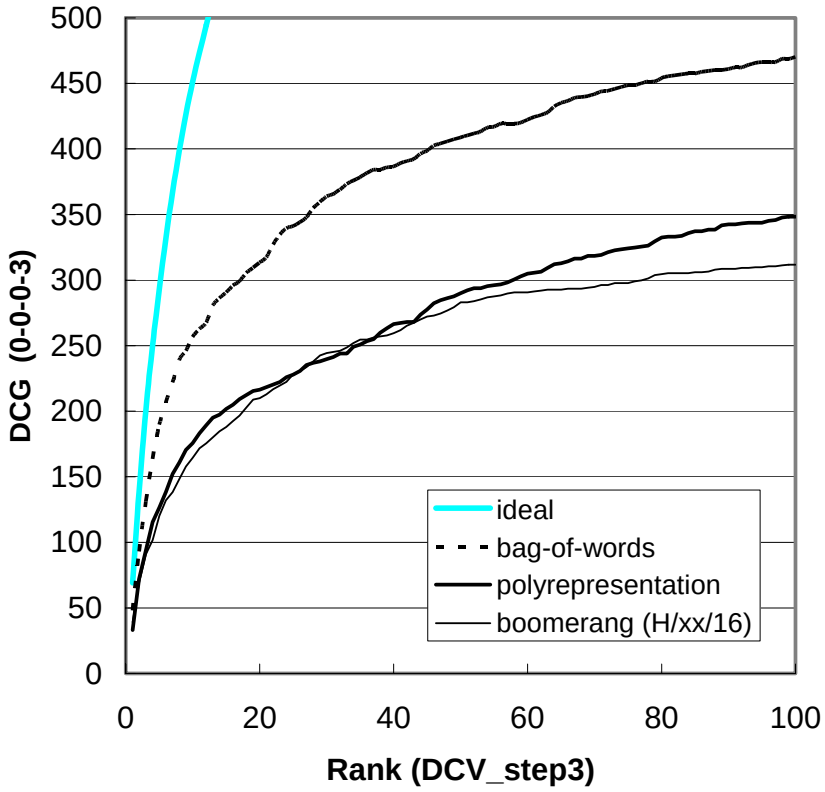



Figure 7.4.b (p. 191):

(nD)CG curves of the best match boomerang effect and the baseline runs as investigated in research question 1. Gain values for the computations were 0-0-0-3, and the natural logarithm (2.718) was used as discount factor for the (n)DCG computations.

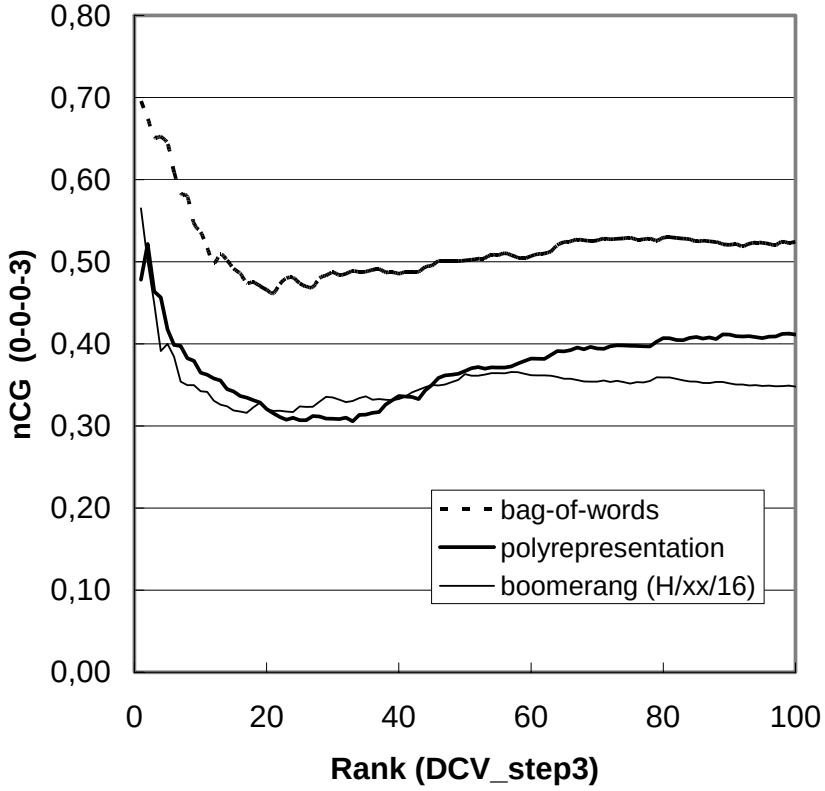



Figure 7.4.c (p. 191):

(nD)CG curves of the best match boomerang effect and the baseline runs as investigated in research question 1. Gain values for the computations were 0-0-0-3, and the natural logarithm (2.718) was used as discount factor for the (n)DCG computations.

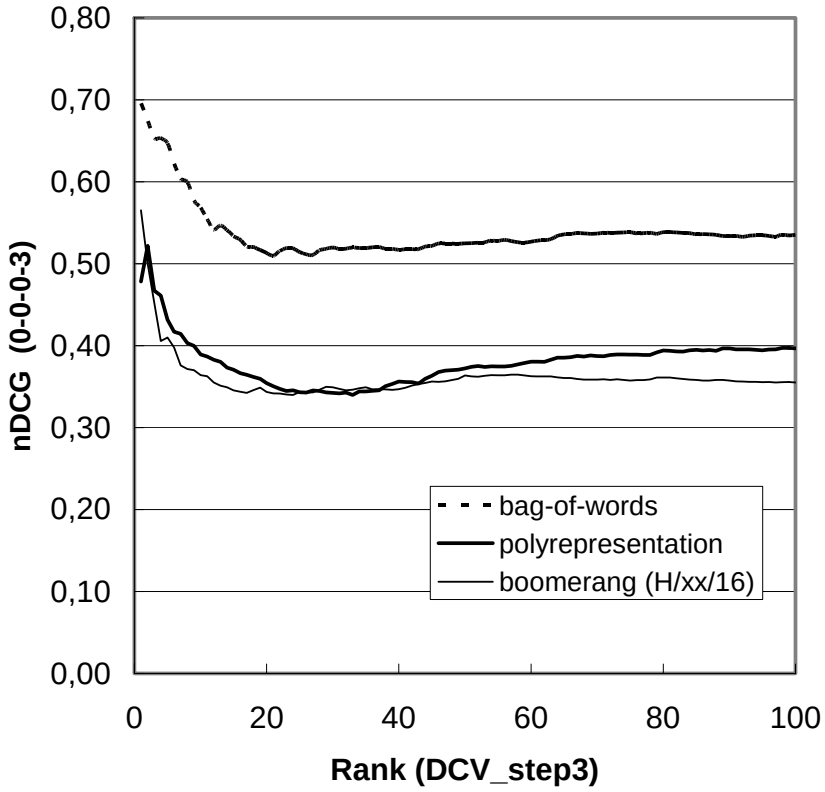



Figure 7.4.d (p. 191):

(nD)CG curves of the best match boomerang effect and the baseline runs as investigated in research question 1. Gain values for the computations were 0-0-0-3, and the natural logarithm (2.718) was used as discount factor for the (n)DCG computations.

# Appendix 10: Examples of CO topics from INEX 2002.

See Section 6.1.2 for details of the markup.

```
<INEX-Topic topic-id="36" query-type="CO">
    <Title>
      <cw>Heat dissipation of microcomputer chips</cw>
    </Title>
    <Description>
      I am looking for generic descriptions on measures taken to combat heat dissipation
      of microcomputer chips.
    </Description>
    <Narrative>
      For a document or document component to be considered relevant, it must give a
      general description on techniques used for 1) lowering energy consumption (by e.g.
      a smarter or more efficient algorithm for some computation), 2) measures taken to
      cool equipment, 3) measures integrated into chip design. The user is not looking
      for mathematical details or models, so document components containing extensive
      mathematical descriptions are not relevant.
    </Narrative>
    <Keywords>
      heat dissipation circuit design heat removal heat generation thermal modelling low
      power
    </Keywords>
</INEX-Topic>

<INEX-Topic topic-id="47" query-type="CO">
    <Title>
      <cw>
        concurrency control semantic transaction management application performance
        benefit
      </cw>
    </Title>
    <Description>
      What are the benefits achieved by deploying semantic transaction management
      techniques.
    </Description>
    <Narrative>
      Relevant documents/components are those that report on performance improvements
      with information systems - especially database systems - when using semantic
      transaction management as opposed to conventional transaction management such as
      two-phase locking. The documents/components should have an analytical
      investigation, a simulation or performance results from a prototype system.
    </Narrative>
    <Keywords>
      "concurrency control" "semantic transaction management" "application" "performance
      benefit" "prototype" "simulation" "analysis"
    </Keywords>
</INEX-Topic>

<INEX-Topic topic-id="58" query-type="CO">
    <Title>
      <cw>Location management scheme</cw>
    </Title>
    <Description>
      Location management scheme for finding mobile object.
    </Description>
    <Narrative>
      A relevant document/component must describe a location management scheme for
      finding a mobile object in the cellular network, wireless LAN or any other
      network. Location management schme may include storing, querying and updating the
      location of an object in the mobile environment.
    </Narrative>
    <Keywords>location management area cell wireless</Keywords>
</INEX-Topic>
```