A Study of the Kinematics of Probabilities in Information Retrieval

by

Fabio A. Crestani

Department of Computing Science
Faculty of Science
University of Glasgow
Glasgow

UNIVERSITY
of
GLASGOW

Thesis submitted for the degree of Doctor of Philosophy

Glasgow, April 1998

"Probability is the very guide of life"

[Cicero, *De Natura*]

Abstract

In Information Retrieval (IR), probabilistic modelling is related to the use of a model that ranks documents in decreasing order of their estimated probability of relevance to a user's information need expressed by a query. In an IR system based on a probabilistic model, the user is guided to examine first the documents that are the most likely to be relevant to his need. If the system performed well, these documents should be at the top of the retrieved list. In mathematical terms the problem consists of estimating the probability $P(R \mid q, d)$, that is the probability of relevance given a query q and a document d. This estimate should be performed for every document in the collection, and documents should then be ranked according to this measure. For this evaluation the system should make use of all the information available in the indexing term space.

This thesis contains a study of the kinematics of probabilities in probabilistic IR. The aim is to get a better insight of the behaviour of the probabilistic models of IR currently in use and to propose new and more effective models by exploiting different kinematics of probabilities. The study is performed both from a theoretical and an experimental point of view.

Theoretically, the thesis explores the use of the probability of a conditional, namely $P(d \rightarrow q)$, to estimate the conditional probability $P(R \mid q, d)$. This is achieved by interpreting the term space in the context of the "possible worlds semantics". Previous approaches in this direction had as their basic assumption the consideration that "a document is a possible world". In this thesis a different approach is adopted, based on the assumption that "a term is a possible world". This approach enables the exploitation of term–term semantic relationships in the term space, estimated using an information theoretic measure. This form of information is rarely used in IR at retrieval time. Two new models of IR are proposed, based on two different way of estimating $P(d \rightarrow q)$ using a logical technique called Imaging. The first model is called Retrieval by Logical Imaging; the second is called Retrieval

by General Logical Imaging, being a generalisation of the first model. The probability kinematics of these two models is compared with that of two other proposed models: the Retrieval by Joint Probability model and the Retrieval by Conditional Probability model. These last two models mimic the probability kinematics of the Vector Space model and of the Probabilistic Retrieval model.

Experimentally, the retrieval effectiveness of the above four models is analysed and compared using five test collections of different sizes and characteristics. The results of this experimentation depend heavily on the choice of term weight and term similarity measures adopted.

The most important conclusion of this thesis is that theoretically a probability transfer that takes into account the semantic similarity between the probability-donor and the probability-recipient is more effective than a probability transfer that does not take that into account. In the context of IR this is equivalent to saying that models that exploit the semantic similarity between terms in the term space at retrieval time are more effective that models that do not do that. Unfortunately, while the experimental investigation carried out using small test collections provide evidence supporting this conclusion, experiments performed using larger test collections do not provide as much supporting evidence (although they do not provide contrasting evidence either). The peculiar characteristics of the term space of different collections play an important role in shaping the effects that different probability kinematics have on the effectiveness of the retrieval process.

The above result suggests the necessity and the usefulness of further investigations into more complex and optimised models of probabilistic IR, where probability kinematics follows non-classical approaches. The models proposed in this thesis are just two such approaches; other ones can be developed using recent results achieved in other fields, such as non-classical logics and belief revision theory.

Acknowledgements

The work reported in this thesis would not have been possible without the help of many people.

First of all, many thanks to Keith van Rijsbergen, my supervisor, for his enthusiastic and constant support throughout my PhD student life. I am very grateful to him for all the freedom he let me enjoy in my work, splitting my time and efforts between Glasgow and Padova.

Thanks to Maristella Agosti for her help and support in the first years of this PhD. She had to make up for my absence from Padova many times.

I am in debt with the people of the Information Retrieval group, in particular with Mark Sanderson, Iain Campbell and Mounia Lalmas for the excellent discussions on every topic somewhat related to Information Retrieval.

Thanks to the all my friends who made and are still making my life in Glasgow enjoyable. I cannot name them all here, but they surely know who I am talking about. They came from many different countries to work or study in Glasgow and most of them are now scattered around the world, but never loose an occasion to get in touch. I learned a lot from them, on topics that have nothing to do with my studies, but that are, by no means, no less important.

Finally, my student life would have been a lot harsher without Monica. She was always there to push me when I lost motivation and to cuddle me when things were not going in the desired direction. I am sure she will be the second happiest person in the world, after me, for the completion of this PhD thesis.

On the financial side, the work reported in this thesis was supported directly by the European Community through the ESPRIT Project FERMI (BRA N. 8134), and indirectly by the University of Padova.

Declaration of Authorship

The material presented in this thesis is the product of my own independent research carried out at the Computing Science Department of the University of Glasgow under the supervision of Professor Cornelis Joost (Keith) van Rijsbergen.

A large amount of the material presented in the thesis has been already published by the author in various technical reports, conference proceedings, and journals articles. In particular, seven chapters of this thesis have been extracted from as many papers, according to the permission granted to me by the Faculty Committee of Higher Degrees of the University of Glasgow on the 27th of November 1997. Some of the papers involved more than one author, however, I hereby declare that I am responsible for the vast majority of the technical/theoretical substance of the papers and of this thesis.

Glasgow, April 1998 Fabio Crestani

Papers Included in the Thesis

The following is a list of the papers used as basis of chapters of this thesis. The list also reports the contributions of the co-authors. The contributions form no more than 5% of the text of this thesis.

Chapter 2 F. Crestani, M. Lalmas, C.J. van Rijsbergen, and I. Campbell. Is this document relevant? ..probably. *ACM Computing Surveys*, in press.

Mounia Lalmas helped me in revising the sections related to "Probabilistic Relevance Models"; Iain Campbell helped in the final revision.

Chapter 3 F. Crestani and C.J. van Rijsbergen. Information Retrieval by Logical Imaging. *Journal of Documentation*, 51(1):1-15, 1995.

Chapter 4 F. Crestani and C.J. van Rijsbergen. A study of probability kinematics in Information Retrieval. *ACM Transactions on Information Systems*, in press.

Chapter 5 F. Crestani, M. Sanderson and C.J. van Rijsbergen. Sense resolution properties of Logical Imaging. *The New Review of Document and Text Management*, 1:277-298, 1995.

Mark Sanderson (University of Glasgow) collaborated with me in writing the sections related to "word sense ambiguity".

Chapter 6 F. Crestani. On the use of Term Space Knowledge for directing the transfer of probabilities in Probabilistic Information Retrieval. *Proceedings of the AIT 96 - Third International Workshop on Artificial Intelligence Techniques*, pages 185-193, Brno, Czech Republic, September 1996.

Chapter 7 F. Crestani and T. Rölleke. Issues on the Implementation of General Imaging on Top of Probabilistic Datalog. In *Proceedings of the First International Workshop on Information Retrieval, Uncertainty, and Logics*, pages 31-41, Glasgow, Scotland, September 1995.

Thomas Rölleke (University of Dortmund) contributed to the sections related to Probabilistic Datalog.

Chapter 8 F. Crestani. Logical Imaging and Probabilistic Information Retrieval. In: F. Crestani, M. Lalmas and C.J. van Rijsbergen, editors, *Logic and Uncertainty in Information Retrieval*, Springer-Verlag, London, UK, in press.

Chapter 9 F. Crestani, I. Ruthven, M. Sanderson, and C.J. van Rijsbergen. The troubles with using a logical model of IR on a large collection of documents. In *Proceedings of the Fourth Text Retrieval Conference (TREC-4)*, pages 509-525, Washington D.C., USA, November 1995.

Ian Ruthven developed the code for the evaluation of EMIM; Mark Sanderson developed the code for performing the experiments.

Prof. C.J. van Rijsbergen, as my Ph.D. supervisor, contributed to a number of the above papers, giving useful suggestions and advices on their technical content and presentation.

Papers Not Included in the Thesis

The following papers of which I am the first author, although containing material related to this thesis, have not been included. Their technical content is either in large part already present in one of the included papers (papers numbered 1, 2, 3, 4, 6), or is the result of a major contribution by the co-author(s) (paper number 5).

1. F. Crestani and C.J. van Rijsbergen. Information Retrieval by Imaging. In *Proceedings of the 16th BCS Colloquium in Information Retrieval*, pages 47-67, Drymen, Scotland, March 1994.

2. F. Crestani and C.J. van Rijsbergen. Probability Kinematics in Information Retrieval: a case study. Technical Report FERMI/95/1, ESPRIT Basic Research Action, Project Number 8134 - FERMI, Glasgow, January 1995.

3. F. Crestani and C.J. van Rijsbergen. Probability Kinematics in Information Retrieval. In *Proceedings of the 18th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 291-299, Seattle, USA, July 1995.

4. F. Crestani. Retrieving Documents by Objective and Subjective General Logical Imaging. In *Proceedings of the Second International Workshop on Information Retrieval, Uncertainty, and Logics*, pages 10-18, Glasgow, Scotland, July 1996.

5. F. Crestani, F. Sebastiani, and C.J. van Rijsbergen. Imaging and Information Retrieval: Variation on a Theme. In *Proceedings of the Second International Workshop on Information Retrieval, Uncertainty, and Logics*, pages 48-49, Glasgow, Scotland, July 1996.

6. F. Crestani. Kinematics of probabilistic term weights and term space knowledge in Information Retrieval. In *Proceedings of the EUFIT 96 - Fourth European Congress on Intelligent Techniques and Soft Computing*, pages 856-905, Aachen, Germany, September 1996.

Contents

List of Figures

List of Tables

Part I

Introduction

Chapter 1

Information Retrieval and Probability

This introductory chapter provides a minimum background knowledge of Information Retrieval necessary to understand the rest of the thesis. The chapter also gives the motivations of the work and outline the structure of the thesis.

1.1 Introduction and motivations

The advent of computers in the last forty years has resulted in an avalanche of machine readable text. Newspapers, books, journals and reports are generated using computers, transmitted by computers, and stored in computers. Bibliographic archives list almost every book, article, and report published. Lawyers have databases with almost every law ever issued and every case ever dealt with. Whole encyclopedias are now published on CD ROMs. Full text of articles from journals are now stored and distributed through the Internet.

Access to information has gone through a slow but steady process to adapt to the growth of availability of electronically stored information. When library were small, access to a piece of information could be achieved by asking the librarian, a "wise sage" who was supposed to have read every book in the library. The librarian could tell you which book contained the information you needed and where the book was located. When the number of books began to exceed the limits of human memory, categorisation became neces-

sary and library classification systems such as the Dewey or the Library of Congress' were developed. Each book was assigned a set of subject headings that identified the topics treated in the book and a location in the library. Only by knowing the appropriate set of subject headings that identified the searched information one could find the location of the book in the library. With computers and the availability of electronic text comes the possibility of searching through the entire text of documents (book, articles, etc.) to find words and phrases that identify a document as containing the information sought. This "free text searching" ability meant that the searcher did not have to rely on someone else looking for documents for him or assigning documents to particular categories. Never the less, if on one hand this puts the searcher in control of the search, on the other hand the searcher now has to know which word to use to express his information need when looking for documents, and every so often he has to know how to use the tool that performs such search.

With the increasing availability of electronic text and with the searcher becoming the user of an information accessing system, it became necessary to develop systems that were both easy to use and effective. New generations Information Retrieval systems need to become easier to user and more effective than current Information Retrieval systems. This thesis will tackle the issue of *effectiveness*.

An Information Retrieval system can become more effective in many different ways, for example by developing a more effective indexing technique to represent better the document informative content, or by capturing better the user information need expressed in the query, or by developing a more effective retrieval technique. This last approach is the one that I tackle in this thesis. I believe that there is ground to develop more effective retrieval techniques, that will make an Information Retrieval system overall more effective, even without improving other components of the system, such as for example indexing or query formulation, although, as every researcher in Information Retrieval recognises, a real breakthrough will only be achieved when new and more advanced document and query representation techniques will be developed. Also, I don't believe in the "pure experimental approach" that many Information Retrieval researchers follow. I think that Information Retrieval has more to gain from an in depth analysis of what has been achieved so far, studying the established results and the pitfalls, than by mere massive runs of experiments blindlessly trying new techniques. An approach mainly focused on experimentation can achieve "ad hoc" improvements of effectiveness, but only by developing new retrieval models based on a deep *theoretical analy-*

sis of the retrieval process will Information Retrieval be able to step firmly forward in the search of more effective systems. This view is also shared by many others Information Retrieval researchers, and in particular by some of the pioneers of this field such as C.J. van Rijsbergen [vR93], S. Robertson [Rob76], and W.S. Cooper [Coo94], whose work greatly inspired me.

One of the most theoretically sound models of Information Retrieval is the *Probabilistic Model.* The Probabilistic Model provides a general theoretical framework for document indexing and retrieval. In this thesis I intend to study possible ways of improving the retrieval process of the Probabilistic Model. To do so I intend to perform a deep analysis of what happens at retrieval time to the probabilities associated to the atomic elements of the probabilistic indexing space. In particular I will study how these probabilities are moved from element to element and how this effects the retrieval performance of a probabilistic Information Retrieval system. I believe that by studying the *kinematics of probabilities* in probabilistic Information Retrieval, we will be able to advance toward developing more effective retrieval technique for next generations of Information Retrieval systems.

1.2 Information Retrieval

Information Retrieval (IR) is the branch of computing science that aims at storing and allowing fast access to a large amount of multimedia information, such as for example text, images, speech, etc. [vR79]. The objects handled by an IR application are usually called *documents*, and the software tool which automatically manages these documents is called *Information Retrieval System* (IRS). The task of an IRS is to help a user to find, in a collection of documents, those documents which contain the information the user is looking for, that is, providing help in satisfying what is often called the *user's information need.*

Frequently IR is confused with database (DB) technology. Figure 1.1 summarises some of the major differences between IR and DB technology; these characteristics have been identified and described by Van Rijsbergen in [vR79], pp. 2. The fundamental difference between IR and DB is that IR systems usually provide only *references* to or a description of the data they manage, while a DBMS provides the actual data. This is not just because of limitations imposed by current technology. Even in full text IR, the task of IR is mainly *to point* at documents. This is because IR systems retrieve documents (mainly) in a probabilistic way, while DB systems retrieve documents

	Information Retrieval	Database
Matching	Partial match	Exact match
Inference	Induction	Deduction
Model	Probabilistic	Deterministic
Classification	Polythetic	Monothetic
Query Language	Natural	Artificial
Query Specification	Incomplete	Complete
Items Wanted	Relevant	Matching
Error Response	Insensitive	Sensitive

Figure 1.1: Differences between IR and DB.

(mainly) in a deterministic way. This means that an IRS retrieves documents that are likely to be considered relevant by the user; that is, likely to satisfy the user's information need. DB facts retrieved in response to a query are always considered to be a complete and true answer to the query, while in IR the perceived relevance of a document varies dramatically across users, and even with one user at different times.

This characteristic of IR has some consequences. First, users' queries to an IRS are usually more vague. They are usually in the form: "I want documents about ...", while users of a DB want facts, such as: "I want the price of the product abc". Second, the evaluation of an IRS is more or less related to its *utility*, that is, how helpful the system is to a user, not a well specified measure, while DBMS are evaluated in accordance with well specified and standardised performance measures.

Given this fundamental differences between IR and DB, an IRS usually manages only descriptions of the informative content of documents. The basic element of these descriptions is called *descriptor* or *index term.* In the classic approach to IR, a schematic view of which is presented in Figure 1.2, the problem of the representation of the document informative content is tackled assigning descriptors to the document. This process is called *indexing*, and it can be either manual or automatic. The representation of the document informative content is one of the most important problems in IR and much efforts are being spent to develop better representations. However, so far the most commonly used indexing technique simply extract descriptors from the text of the document performing a quite simple lexical analysis.

Once a suitable representation structure has been provided, an IRS faces the problem of evaluating the similarity between query and document representations. This is often achieved by evaluating a *similarity measure* which uses

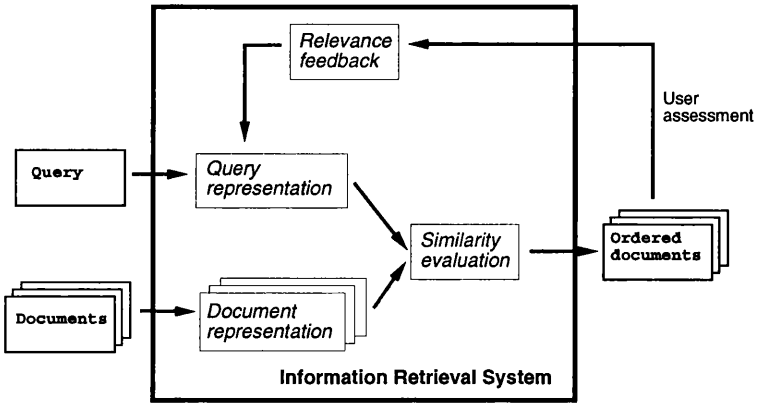

Figure 1.2: A schematic view of a classical Information Retrieval system.

features of document and query representation to evaluate an overall degree of similarity between the query and each document in the collection.

An advance technique becoming rapidly available on most IRS is *relevance feedback*. Relevance feedback is a technique that allows a user to express in a better way his information requirement by modifying his original query formulation with further information provided by indicating some relevant documents [Har92b]. When a document is marked as relevant the relevance feedback algorithm analyses the document text, picking out terms that are statistically significant, and adds these terms to the query. Relevance feedback is a very good technique for specifying an information requirement, because it releases the user from the burden of having to think up lots of terms for the query. Instead the user deals with the ideas and concepts contained in the documents. It also fits in well with the known human trait of "I don't know what I want, but I'll know it when I see it". Obviously the user cannot mark documents as relevant until some are retrieved, so the first search has to be initiated by a query. In response to this initial query, the IRS will return a list of ordered documents covering a range of topics, but probably at least one document in the list will cover, or come close to covering, the user's interest. The user marks some document(s) as relevant and starts the relevance feedback process performing another search. If the relevance feedback algorithm performs well the next list should contain documents closer to the user's requirement, and the process can be repeated until the user is satisfied by the result.

In the next section I will provide a brief overview of the state of the art of IR and of the evaluation techniques used for measuring the effectiveness of IR systems. This will provide a general framework in which to place the

contribution of this thesis.

1.3 The state of the art of Information Retrieval

IR is an established technology that has been delivering solutions to users for more than 30 years and yet it is still an active area of research. This suggests that although much work has been done, much remains to be accomplished. In over 30 years, researchers in IR have developed and evaluated a bewildering array of techniques for indexing and retrieving text. These techniques have slowly matured and improved through refinement rather than there having been one or a small number of really significant breakthroughs.

In this section I will give a brief report of the state of the art of IR, showing how we can see a clear distinction between the systems used in the research world (experimental systems), the brand new class of IR systems that is concerned with searching the World Wide Web (WWW), and the systems used in the commercial world (libraries, information providers, etc.).

1.3.1 Experimental IR Systems

Experimental IR systems are systems that have been developed to test new indexing or retrieval techniques. They are the means by which IR researchers test new ideas. An experimental IR system is often not very efficient with respect to CPU time and memory resources. The main purpose of the system is to enable a fast testing of some theoretical ideas, therefore the development of routines for the efficient use of computer resources is left for a later time.

Experimental systems have enabled IR researchers to develop and evaluate many new indexing and retrieval techniques [SJ81] and have been and still are very important vehicles of research. The SMART system of Cornell University, for example, was developed almost 30 years ago as an experimental IR system for testing a particular IR model (the Vector Space model) and is still a very important instrument of research [Sal71]. There are many other examples of experimental systems, MEDLARS [Mil71], STAIRS [Bla96], PThomas [OB91], I^3R [CR87], RIME [BC89], GRANT [KC87], just to mention a few.

In the context of this thesis I will use an experimental system called SIRE,

developed at Glasgow University. SIRE will be used as an implementation platform for testing the theoretical ideas proposed here.

Some indexing and retrieval techniques developed and tested on experimental systems are later exploited commercially, while others never make it to be used in commercial systems. There are many reasons for this. The most common reasons are that the technique developed is highly inefficient (this is perhaps the most important reason) or not effective enough to justify the investment. While the efficiency problem can be tackled or perhaps simply solved by the availability of faster computers, the effectiveness problem has no simple solution and requires a re-thinking of the proposed techniques.

1.3.2 Web Search Engines

The *World Wide Web* (WWW, W3 or Web) is a recent but explosive phenomenon [BL96]. The Web is a global information system that provides hypertext access to resources on the Internet via a common syntax of addressing network resources (URL), a common protocol for the transfer of data from a Web server to a Web client (HTTP), and a mark-up language (HTML) for writing the hypertext nodes. The Web client is responsible for communicating with servers to retrieve necessary documents and files. The Web server is responsible for making local documents or files available to other software systems.

The number of available sources of information on the Web has constantly increased over the last few years: during 1993 the number of Web sites has passed from 50 to 500, but by the end of 1995 the sites world-wide available were more than 10,000. The Web also incorporates existing network services, such as FTP and Gopher. Because of this expansion of Web sites, to find specific information on some topic using only the browsing paradigm over the Web is almost impossible.

Documents located on a single site can often be searched by ignoring the link structure and applying full text search like with text archives. But searching of information on a specific topic over the whole Web would not be possible without the help of one of the search tools that have been rapidly developed and made available on the Web.

Web search tools are very useful tools that enable a user to search for navigation starting points of the Web that satisfy his query. Web search tools often use technology developed in the IR area. Sometimes it is consolidated

technology, like in the case of most large commercial tools, but some other times Web search tools use techniques just out of experimental IR systems. The testing (for free) of these techniques on such a large document set and with a large number of users provides a very useful assessment of their effectiveness. In this case Web search tools can be considered the intermediate step for a techniques developed at experimental level before being used in commercial systems.

There are several Web search tools, differing substantially by many characteristics. A tentative classification of these search tools can be organised as follows:

- *Search engines*: tools for finding documents on the Web based on their contents (e.g. Lycos, Altavista).
- *Meta-search engines*: search engines that consults several search engines at the same time (e.g. SavvySearch, Find-It!).
- *Subject directories*: tools to search Web sites that base the search on what the site is about (e.g Yahoo).
- *Geographical directories*: tools that makes a search on Web sites, but the search is based on where the site is located (e.g. The Virtual Tourist).
- *Link databases*: tools for finding links in databases of Web sites and resources (e.g. IWeb).

Other search tools have been developed and made available to search for information all over Internet and not only in Web sites. Some of these are: people directories and white pages, business directories and yellow pages, software archives, newsgroups and mailing lists search, background information search.

From the point of view of the work reported in this thesis, among the above list of search tools, the search engines are the most interesting. A search engine is a networked IRS that operates over the Web. The main difference between a search engine and an IRS is that a search engine operates in a sort of symbiosis with a Web robot. The robot is used to index the content of Web pages and to follow the links that are present in the Web pages to collect and find further information. The robot can be classified as an automatic browser that is able to autonomously traverse the complete Web

structure finding each single Web document and recursively finding all the other documents that are related to it through Web links. Most common synonyms of Web robot are: crawler, wanderer, spider, and worm. Web ants are robots that work in a co-operative way and in parallel to save time [Men95].

Whatever the way a search engine uses to index and represent document out of the Web, it still has to rely on a similarity algorithm to evaluate how documents match queries. From this point of view a search engine is not different from an IRS, and the results of this thesis could be used to build new Web search engines too.

1.3.3 Commercial IR Systems

Techniques that are developed and tested in experimental systems take a long time before being accepted to be included in commercial IR systems. Most commercial IR system are still based on the Boolean model, one of the oldest IR model, and only recently systems based on free-text queries have started to be accepted in the commercial world. There are many reasons why this happens. Perhaps the most important one is that most users of commercial IR systems are middle-aged professionals that received their training on old commercial systems based on the Boolean model, such as STAIRS [Bla96] and MEDLINE [Fei85], for example.

However, things are changing. First of all, the newer generation of professional users of commercial systems have been trained not only on Boolean systems, but also on free-text systems. They feel equally comfortable using both models. Second, but perhaps more important, IR systems are more and more being accessed by end users without the intervention of professional intermediaries. These end users, most of them not expert in IR, want IR systems that are easy to use and that do not require a long training to be used effectively. Interactive free-text partial-match IR systems seem to respond to these users needs. It is towards this type of systems that the work reported in this thesis is directed.

1.3.4 Evaluation of IR Systems

Much effort and research has gone into studying the problem of evaluation in IR. Never the less, most of the people active in this field still feel that the

documents:	*relevant*	*not relevant*	
retrieved	$A \cap B$	$\neg A \cap B$	B
not retrieved	$A \cap \neg B$	$\neg A \cap \neg B$	$\neg B$
	A	$\neg A$	

Figure 1.3: Determination of precision and recall values

problem is far from solved, in particular for interactive IR systems. Here I will not enter into the discussion about which evaluation technique is best, but I will simply report on the "state of the art" of the IR evaluation techniques.

Following the approach proposed by Van Rijsbergen in [vR79], in trying to evaluate an IRS one has to try to answer to at least two questions:

1. what to evaluate?
2. how to evaluate?

The answer to the first question is related to the main purpose of one's work. A researcher could be interested in evaluating the speed of some retrieval process, or the level of interaction an IRS allows, for example. There are various aspects of IR that a researcher could be interested in evaluating. I will not address the general issues of what is the most important feature of an IRS to evaluate. The main feature I will consider in the evaluation is the *effectiveness* of retrieval. In doing this I am in tune with the most part of the evaluations reported in the IR literature.

The second question instead needs a more technical answer. Among the various measurable quantities of effectiveness proposed as early as in 1966 by Claverdon [CMK66], the *time lag*, the *presentation*, and the *effort*, have been almost completely ignored in the IR literature, because they are related to an operational implementation of the system and not to the effectiveness of the IR process. The two best known measures of effectiveness are *recall* and *precision*. They are by far the measures of effectiveness most commonly used in the IR literature. Since these two measures will be used extensively to report the results of the evaluations described in this thesis, I will explain them in detail.

In order to have clear the meaning of the recall and precision measures, their definition, as described in [vR79], is here reported. It is helpful to refer to Figure 1.3, from which recall and precision can easily be derived. They are

defined as:

$$Precision = \frac{| A \cap B |}{| B |}$$

$$Recall = \frac{| A \cap B |}{| A |}$$

where $| A \cap B |$ is the number of relevant and retrieved documents, $| B |$ is the number of retrieved documents, and $| A |$ is the number of relevant documents.

The evaluation of these precision and recall values is only possible if one has complete knowledge of the relevant documents present in the collection. This is not possible in most operative cases. Therefore, in order to enable an evaluation of IR systems (in particular experimental IR systems) a considerable amount of resources have been spent in building *test collections* [Sv76]. These are collections of textual documents that come with a set of queries and lists of documents in the collection that are known to be relevant to the queries. The availability of relevance judgements enables the evaluation of precision and recall values for diverse indexing and retrieval strategies in a controlled environment. This makes it possible to compare the results and draw conclusions that can be extended to operative cases.

In the evaluation of an IR system, precision and recall values need to be evaluated for every query submitted to the system. However, the evaluation of these values depends on cut-off points in the ranked list of documents retrieved in response to the query. Therefore, a better way of displaying these measures is through a *recall-precision graph*. An example of such a graph is depicted in Figure 1.4, where precision values are reported corresponding to standard recall values.

To measure the overall performance of the system on a set of queries, it is necessary to produce as many graphs as the number of queries and then combine them in some way. This is often done using the "macro-evaluation" approach, which consists in averaging over all queries the individual precision values corresponding to the standard recall values. All the graphs reported in this thesis are obtained in this way. For a more in depth explanation of IR evaluation techniques see chapter seven of Van Rijsbergen's book [vR79].

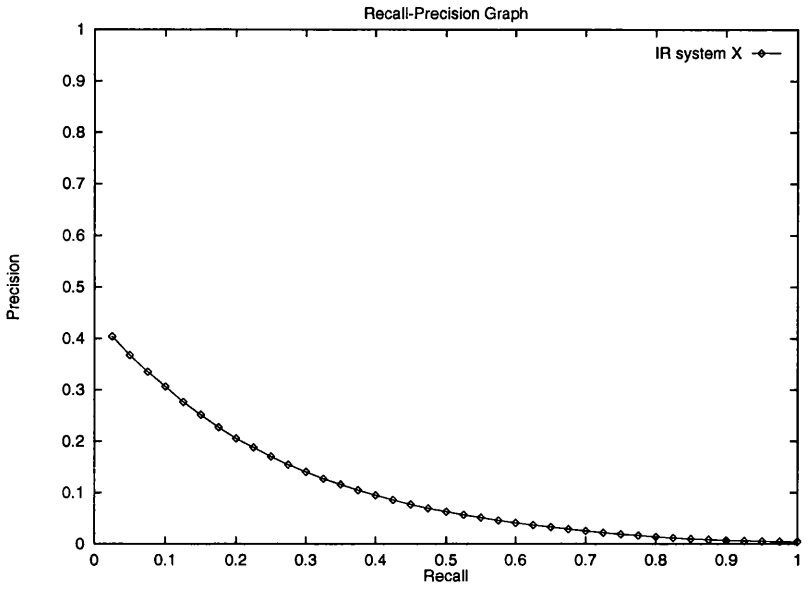

Figure 1.4: An example of a Recall-Precision graph

1.4 Probabilistic Information Retrieval

The basic belief of *probabilistic approaches to IR* is that, for optimal performance, documents should be ranked in order of decreasing probability of relevance or usefulness to the user. Probabilistic approaches therefore attempt to estimate or calculate in some way the probability that a document will be relevant to a particular user need expressed in a natural language query. The explicit formulation of this idea is given in the *The Probability Ranking Principle* [Rob77], which states:

> "If a reference retrieval system's response to each request is a ranking of the documents in the collection in order of decreasing probability of usefulness to the user who submitted the request, where the probabilities are estimated as accurately as possible on the basis of whatever data has been made available to the system for this purpose, then the overall effectiveness of the system to its users will be the best that is obtainable on the basis of that data."

However, the interpretation of the phrase probability of relevance is far from straightforward and a number of different connotations have been put upon

it. Moreover, the evaluation of the probability of relevance from the available data is far from simple and many different models have been proposed for this purpose. I will review the various approaches to probabilistic IR in Chapter 2.

Although there are a few experimental IR systems based on probabilistic or semi-probabilistic models, there are still obstacles to getting probabilistic models accepted in the commercial IR world. One major obstacle is that of finding methods for estimating the probabilities of relevance that are both effective and computationally efficient. Past and present research has made much use of formal probability theory and statistics in order to solve the problems of estimation. In mathematical terms the problem consists in estimating the probability $P(R \mid q, d)$, i.e. the probability of relevance given a query q and a document d for every document in the collection, and ranking the documents according to this measure. This is very difficult because of the large number of variables involved in the representation of documents in comparison with the small amount of feedback data available about the relevance of documents, a problem sometimes refered to as the "curse of dimensionality" [Eft96].

In 1986 Van Rijsbergen [vR86] proposed to consider the conditional probability $P(R \mid q, d)$ as the probability of the conditional $d \rightarrow q$, that is $P(d \rightarrow q)$. In order to evaluate $P(d \rightarrow q)$ he proposed to use the following *logical uncertainty principle*:

> "Given any two sentences x and y; a measure of the uncertainty of $y \rightarrow x$ related to a given data set is determined by the minimal extent to which we have to add information to the data set, to establish the truth of $y \rightarrow x$."

However, Van Rijsbergen said nothing about how "uncertainty" and "minimal" might be quantified.

The logical uncertainty principle initiated a new line of research that has been followed by many researchers, (see for example [Nie88, CC92, Bru93]), and different interpretations of the term "uncertainty" and different ways to estimate it have been proposed.

1.5 Probability Kinematics and Probabilistic Information Retrieval

The logical uncertainty principle is the staring point of this thesis. However, to proceed with its application we need to map it first to the IR problem. To do so we need to answer the following questions in the IR context:

1. What are x and y?
2. How do we interpret $y \rightarrow x$? What is the semantics of $y \rightarrow x$?
3. How do we choose $\mu(y \rightarrow x)$, the measure of the uncertainty of $y \rightarrow x$?
4. What is the data set in the context of which $\mu(y \rightarrow x)$ is evaluated?
5. How can we add information to the data set? Where does this information come from?
6. How do we establish the truth of $y \rightarrow x$?
7. How can we measure the information we have added to the data set to establish the truth of $y \rightarrow x$? How do we calculate $\mu(y \rightarrow x)$?
8. How can we be sure that the information we added to the data set is minimal?

The rest of the thesis will be devoted to provide answers and explain solutions to the above questions. The result is a set of models for probabilistic IR that are based on the logical uncertainty principle. Note that this thesis tackles these questions in a very particular way, making decisions that may be questionable, but that are reasonable and theoretically sound.

Two *assumptions* have been taken in order to answer the above questions:

- The measure of uncertainty μ will be searched in the context of Probability Theory, so that the resulting model will be a probabilistic model of IR.
- Once we have specified a probabilistic space, the information to be added to the data set to establish the truth of $y \rightarrow x$ will come from elements of the probabilistic space itself and not from outside it.

The above two assumptions were taken with the purpose of restricting the area of research, so providing a framework inside which carry out an in depth research. With the first assumption I restrict the choice of μ into a very well delimited framework. The *probabilistic framework* has proved to be very successful in IR, although other approaches based on different uncertainty measures have been proposed (see for example [Lal96, TC92b]). The second assumption assures that I will not be concerned with information provided by a user, an intermediary, or some other external source. The introduction of the user in this area of research would generate a number of issues related to the subjectivity of the results, the interaction with the user, the modelling of the user behaviour, and so on. I do not wish to tackle these issues at this stage.

Given the above two assumptions, this thesis will be concerned with specifying a probabilistic space in which the basic elements of IR, that is index terms, documents, and queries, will be placed and assigned probabilities to. These probabilities will then be moved around in the retrieval space at retrieval time in order to achieve the goal of a probabilistic IR system, that is to rank documents in decreasing order of some estimated probability of relevance to a user query. Existing probabilistic IR models will also be studied in this same probabilistic space, to compare their behaviour with that of the new models I will propose.

Particular emphasis will be placed to the *kinematics of probabilities*, that is to the study of how probabilities are moved from one element of the probabilistic indexing space to another. I believe that the key to obtain better probabilistic IR models, both from a theoretical and a practical point of view, is to study existing models at a very deep level, almost like studying them "through a microscope". This will enable a more in depth comparison with existing models and a more accurate explanation of the differences in behaviour and results.

1.6 Structure of the thesis

A large amount of the technical material reported in the rest of this thesis has been already published by me in various technical reports, workshops and conference proceedings, and journals articles. In particular, eight chapters of the thesis (from Chapters 2 to 9 included) have been extracted from papers either published or in the process of being published. A detailed list of where and when these papers have been published is reported at the beginning

of this thesis in the Declaration section. The work reported here forms a seamless study on the kinematics of probabilities in probabilistic IR.

In particular, Chapter 2 reports a survey of the *state of the art* of probabilistic modelling in IR. This provides the background in which the study reported in this thesis should be placed. Chapter 3 presents a first formulation of a new class of probabilistic retrieval models based on Logical Imaging. This formulation is later enhanced and compared with probabilistic models of IR in Chapter 4. Chapter 5 reports an analysis of the sense resolution properties of the proposed models. Chapter 6 tackles the problem of using these new models in the absence of complete information on the probabilistic term space. These last four chapters constitute a *theoretical study* that is the core of the thesis. The following two chapters report on the *implementation study* into possible implementation platforms for the proposed models. Chapter 7 reports on the implementation of the proposed models on top of Probabilistic Datalog, while Chapter 8 reports on their implementation on top of the $\mathcal{L}_1$ Probabilistic Logic. The following three chapters report on the *experimental study* into the effectiveness of the proposed models. Chapter 9 reports on the troubles faced in experimenting retrieval by Logical Imaging with a very large collection of documents. Chapter 10 reports on further experimentations with a large collection of documents and analyses some contrasting experimental results obtained from different implementations of retrieval by Logical Imaging and General Logical Imaging. Chapter 11 concludes the thesis reporting the theoretical and the experimental conclusions, and the future work.

Note on the Compilation of this Thesis

It is general practice when writing research papers to introduce the topic of research by recalling theoretical points and results already presented by the author in previous papers. For this reason, some concepts that are central to this thesis are reported a few times in the chapters derived from papers. Although this may seem a useless repetition, it is a consequence of the original structure of the papers from which this thesis has be derived and could not be totally avoided.

Part II

State of the Art

Chapter 2

Probabilistic Information Retrieval

This chapter provides an introduction to and survey of probabilistic approaches to modelling Information Retrieval. The basic concepts of probabilistic approaches to Information Retrieval are outlined, and the principles and assumptions upon which the approaches are based are presented. The various models that have been proposed in the development of IR are described, classified, and compared. The models are classified and compared using a common formalism. New approaches that constitute the basis of future research are described.

2.1 History of probabilistic modelling in IR

In Information Retrieval (IR), probabilistic modelling is the use of a model that ranks documents in decreasing order of their evaluated probability of relevance to a user's information need. Past and present research has made much use of formal theories of probability and of statistics in order to evaluate, or at least estimate, those probabilities of relevance. These attempts are to be distinguished from looser ones like, for example, the Vector Space model [Sal68] in which documents are ranked according to a measure of similarity with the query. A measure of similarity cannot be directly interpretable as a probability. In addition, similarity based models generally lack the theoretical soundness of probabilistic models.

The first attempts to develop a probabilistic theory of retrieval were made

over thirty years ago [MK60, Mil71]. Since, there has been a steady development of the approach. There are already several operational IR systems based upon probabilistic or semi-probabilistic models.

One major obstacle with probabilistic or semi-probabilistic IR models is that of finding methods for estimating the probabilities used to evaluate the probability of relevance that are both theoretically sound and computationally efficient. The problem of estimating these probabilities is difficult to tackle unless some simplifying assumptions are made. In the early stages of the study of probabilistic modelling in IR, assumptions related to event independence were employed in order to facilitate the computations. The first models to be based upon such assumptions were the "binary independence indexing model" (Section 2.3.3) and the "binary independence retrieval model" (Section 2.3.2). Recent findings by Cooper [Coo95] have shown that these assumptions are not completely necessary and were, in fact, not actually made (Section 2.5).

The earliest techniques that took into account dependencies gave results that were worse that those given by techniques based upon the simplifying assumptions. Moreover, the use of complex techniques that captured dependencies could only be made at a computational price regarded as too high with respect to the value of the results [vR77]. One particular research direction aimed at removing the simplifying assumptions has been studied extensively and much work is being done [FCAT90, TC90, Sav92, vR92].

Another direction has involved the application of the statistical techniques used by pattern recognition and regression analysis. These investigations, of which the "Darmstadt indexing approach (DIA)" is a major example [Fuh89, FB91] (see Section 2.3.4), do not make use of independence assumptions. They are "model free" in the sense that the only probabilistic assumptions involved are those implicit in the statistical regression theory itself. The major drawback of such approaches is the degree to which heuristics are necessary to optimise the description and retrieval functions.

A theoretical improvement of the DIA was achieved through the use of logistic regression instead of standard regression. Standard regression is, strictly speaking, inappropriate for estimating probabilities of relevance where relevance is considered as a dichotomous event: i.e. a document is either relevant to a query or not. Logistic regression has been specifically developed to deal with dichotomous (or n-dichotomous) dependent variables. Probabilistic models that make use of logistic regression have been developed by Fuhr and Pfeifer in [FB91] and by Cooper et al. in [CGD92] (Sections 2.3.4 and

2.3.7).

One area of recent research investigates the use of an explicit network representation of dependencies. The networks are processed by means of Bayesian inference or belief theory, using evidential reasoning techniques such as those described by Pearl [Pea88]. This approach represents an extension of the earliest probabilistic models, taking into account the conditional dependencies present in a real environment. Moreover, the use of such networks generalises existing probabilistic models and allows the integration of several sources of evidence within a single framework. Attempts to use Bayesian (or causal) networks are reported in [Tur90, TC91, Sav92].

There is also a new stream of research, initiated by van Rijsbergen [vR86] and continued by him and others [AvR95, Lal96, Bru93, Bv92, Hui96, Seb94, Cv95]. It aims at developing a model based upon a non-classical logic, in particular, a conditional logic where the semantics is expressed using probability theory. The evaluation can be performed by means of a possible-world semantics [vR89, vR92, Sv93, CvR95] thus establishing an intentional logic, by using modal logic [Nie88, Nie89, AK92, Nie92], by using situation theory [Lal92], or by integrating logic with Natural Language Processing [CC92]. The area is in its infancy; no working prototype based on the proposed models has been developed so far, and the operational validity of these ideas has still to be confirmed.

2.2 Background

In this section, I review some general aspects that are important for a full understanding of probabilistic models. Then, I provide a framework within which the various models can be placed for comparison. I do not deal with concepts of probability theory in this chapter. I assume some familiarity of principles of probability theory on the part of the reader. Finally, because of its importance to the foundations of all probabilistic retrieval models, I present the Probability Ranking Principle.

2.2.1 Event space

In general, probabilistic models have as their event space the set $\underline{Q}$ x $\underline{D}$, where $\underline{Q}$ represents the set of all possible queries, and $\underline{D}$ the set of all documents

in the collection. The difference between the various models lies in their use of different *representations* and *descriptions* of queries and documents.

In most models, queries and documents are represented by descriptors, often automatically extracted or manually assigned terms. These descriptors are represented as binary valued vectors in which each element corresponds to a term. More complex models make use of real valued vectors, or take into account relationships among terms or among documents.

A *query* is an expression of an information need. In this thesis, I regard a query as a unique event; that is, if two users submit the same query, or if the same query is submitted by the same user on two different occasions, these two queries are regarded as different queries. A query is submitted to the system, which then aims to find information *relevant* to the information need expressed in the query. In this thesis I will consider relevance as a subjective user judgement on a document related to a unique expression of an information need[1].

A *document* is any object carrying information; a piece of text, an image, a sound, or a video. However, most current IR systems deal only with text. This limitation results from problems associated with finding suitable representations for non textual objects. Therefore, in the remaining of this thesis, I will consider only text-based IR systems.

Some assumptions that are common to all retrieval models:

- The users' understanding of their information need changes during a search session, is subject to a continuous refinement, and is expressed by different queries.
- Retrieval is based only upon representations of queries and documents, not upon the queries and documents themselves.
- The representation of IR objects is "uncertain". For example, the extraction of index terms from a document or a query to represent the document or query informative content is a highly uncertain process. As a consequence, the retrieval process becomes uncertain.

[1]There exists a relevance relationship between a query and a document, which relies on a user perceived satisfaction of his or her information need. Such a perception of satisfaction is subjective - different users can give different relevance judgements to a given query-document pair. Moreover, this relevance relationship depends on time, so the same user could give a different relevance judgement on the same query-document pair on two different occasions.

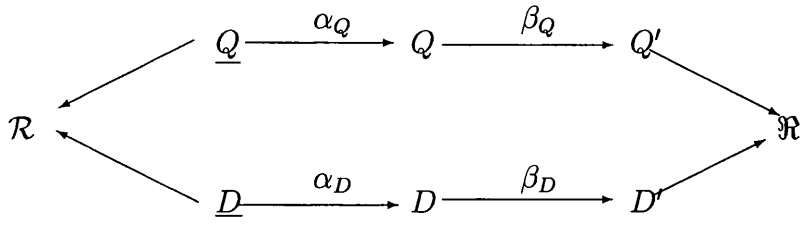

Figure 2.1: The underlying conceptual model.

It is particularly this last assumption that gave way to the study of probabilistic retrieval models. Probability theory [Goo50] is, however, only one way of dealing with uncertainty[2]. Also, earlier models were largely based on classical probability theory, but, in recent times new approaches to dealing with uncertainty have been applied to IR. Sections 2.3 and 2.4 present both traditional and new approaches to probabilistic retrieval.

2.2.2 A conceptual model

The importance of conceptual modelling is widely recognised in fields such as Database Management Systems and Information Systems. For this thesis, I will use the conceptual model proposed by Fuhr [Fuh92b], which has the advantage of being both simple and general enough to be considered a conceptual basis for all probabilistic models presented in this survey, although some of them predate it.

The model is shown in Figure 2.1. The basic objects of an IR system are: a finite set of *documents* $\underline{D}$ (e.g., books, articles, images) and a finite set of *queries* $\underline{Q}$ (e.g., information needs). I consider a set of queries and not a single query alone because a single user may have varying information needs. If we consider $\mathcal{R}$ a finite set of possible relevance judgements, for example in the binary case $\mathcal{R} = \{R, \overline{R}\}$, that is, a document can either be relevant or not to a query, then the IR system's task is to map every query-document pair to an element of $\mathcal{R}$. Unfortunately, IR systems do not deal directly with queries and documents, but with *representations* of them (e.g., a text for a

[2]Other approaches are based, for example, on Fuzzy Logic [Zad87] and Dempster-Shafer's theory of evidence [Sha76].

document, or a Boolean expression for a query). It is mainly the kind of representation technique used that differentiates one IR model from another.

I denote α_Q the mapping between a set of queries $\underline{Q}$ and their representations Q. For example, a user in search of information about wine may express his or her query as follows: "I am looking for articles dealing with wine". Similarly, I denote α_D the mapping between a set of documents $\underline{D}$ and their representations D. For example, in a library, a book is represented by its author, titles, a summary, the fact it is a book (and not a article), and some keywords. These two mappings can be very different from each other. Obviously, the better the representation of queries and documents, the better will be the performance of the IR system.

To make the conceptual model general enough to deal with the most complex IR models, a further mapping has been introduced between representations and *descriptions*. For instance, a description of the above query could be the following two stems: "article" and "wine". The sets of representations Q and D are mapped to the sets of descriptions Q' and D' by means of two mapping functions β_Q and β_D. Moreover, the need for such additional mapping arises for learning models (see for example Section 2.3.4) that have to aggregate features to allow large enough samples for estimation. It is worth noticing, however, that most models work directly with the original document and query representations.

It is common for IR systems to be able to manage only a poor description of the representation of the objects (e.g., a set of stems instead of a text). However, when representation and description happen to be the same, it is sufficient to consider either α_Q or α_D as an identity mapping.

Descriptions are taken as the independent variables of the retrieval function $r : Q' \times D' \rightarrow \Re$, which maps query-document pair onto a set of *retrieval status values* (RSV) $r(q'_k, d'_j)$ [BC76]. The task of ranked retrieval IR systems in response to a query $\underline{q}_k$ is to calculate this value and rank each and every document $\underline{d}_j$ in the collection upon it.

In probabilistic IR the task of the system is different. If we assume binary relevance judgements, i.e. $\mathcal{R}$ contains only the two possible judgements R and $\overline{R}$, then according to the Probability Ranking Principle (Section 2.2.4), the task of an IR system is to rank the documents according to their estimated probability of being relevant $P(R \mid \underline{q}_k, \underline{d}_j)$. This probability is estimated by $P(R \mid q'_k, d'_j)$, which is the retrieval status value.

2.2.3 On the concepts of "relevance" and "probability of relevance"

The concept of *relevance* is arguably the fundamental concept of IR. In the above presented model I purposely avoid giving a formal definition of relevance. The reason behind this decision is that the notion of relevance has never been defined precisely in IR. Although there has been a large number of attempts towards a definition of the concept of relevance [Ser70, Coo71, Miz96], there has never been agreement about unique precise definition. A treatment of the concept of relevance is outside the scope of this chapter and I will not attempt to formulate a new definition or even accept a particular already existing one. What is important for the purpose of this survey is to understand that relevance is a relationship that may or may not hold between a document and a user of the IR system who is searching for some information: if the user wants the document in question, then we say that the relationship holds. With reference to the model presented above, relevance ($\mathcal{R}$) is a relationship between a document ($\underline{d_j}$) and a user's information need ($\underline{q_k}$). If the user wants the document $\underline{d}$ in relation to his information need $\underline{q_k}$, then $\underline{d_j}$ is relevant (R).

Most readers will find the concept of *probability of relevance* quite unusual. The necessity of introducing such probability arises from the fact that relevance is a function of a large number of variables concerning the document, the user, and the information need. It is virtually impossible to make strict prediction as to whether the relationship of relevance will hold between a given document and a given user's information need. The problem must be approached probabilistically. The above model explains what is the evidence available to an IR system to estimate the probability of relevance $P(R \mid \underline{q_k}, \underline{d_j})$. A precise definition of probability of relevance depends on a precise definition of the concept of relevance, and given a precise definition of relevance it is possible to define rigorously such probability. Just as I did not define relevance, I will not attempt to define the probability of relevance, since every model presented here uses a somewhat different definition. I refer the reader to the treatment given by Robertson et al. in [RMC82], where different interpretations of the probability of relevance are given and a unified view is proposed.

2.2.4 The Probability Ranking Principle

A common characteristic of all the probabilistic models developed in IR is their adherence to the theoretical justification embodied in the Probability Ranking Principle (PRP) [Rob77]. The PRP asserts that optimal retrieval performance can be achieved when documents are ranked according to their probabilities of being judged relevant to a query. The above probabilities should be estimated as accurately as possible on the basis of whatever data has been made available for this purpose.

The principle speaks of "optimal retrieval", as distinct from "perfect retrieval". Optimal retrieval can be defined precisely for probabilistic IR because it can be proved theoretically with respect to representations (or descriptions) of documents and information needs. Perfect retrieval relates to the objects of the IR systems themselves, i.e., documents and information needs.

The formal definition of the PRP is as follows. Let C denote the cost of retrieving a relevant document, and $\overline{C}$ the cost of retrieving a non-relevant document. The decision rule that is the basis of the PRP states that a document d_m should be retrieved in response to a query q_k above any document d_i in the collection if:

$$C \cdot P(R \mid q_k, d_m) + \overline{C} \cdot (1 - P(R \mid q_k, d_m)) \leq P(R \mid q_k, d_i) + \\ + \overline{C} \cdot (1 - P(R \mid q_k, d_i))$$

The decision rule can be extended to deal with multi-valued relevance scales, (e.g., very relevant, possibly relevant, etc. [Coo71]). In addition, by means of a continuous cost function, it is possible to write a decision rule for approaches where the relevance scale is assumed to be continuous [BP93].

The application of the PRP in probabilistic models involves assumptions:

- Dependencies between documents are generally ignored. Documents are considered in isolation, so that the relevance of one document to a query is considered independent from that of other documents in the collection (nevertheless, see Section 2.5).

- It is assumed that the probabilities (e.g., $P(R \mid q_k, d_i)$) in the decision function can be estimated in the best possible way, that is accurately enough to approximate the user's real relevance judgement, and therefore order the documents accordingly.

Although these assumptions limit the applicability of the PRP, models based on it enable the implementation of IR systems offering some of highest level of retrieval performance currently available [Rob77]. There are, of course, a number of other retrieval strategies with high levels of performance and that are not consistent with the PRP. Examples of such strategies are the Boolean or the cluster model. In this chapter I am not concerned with these models since they are not probabilistic in nature and do not fall into the class of models this survey is about.

2.2.5 The remainder of this chapter

In the remainder of this chapter, I present a survey of probabilistic IR models in two main categories: *relevance models* and *inference models*.

Relevance models are described in Section 2.3. These models are based on evidence about which documents are relevant to a given query. The problem of estimating the probability of relevance for every document in the collection is difficult because of the large number of variables involved in the representation of documents in comparison to the small amount of document relevance information available. The models differ, primarily, in the way they estimate this or related probabilities.

Inference models are presented in Section 2.4. These models apply concepts and techniques originating from areas such as logic and artificial intelligence. From a probabilistic perspective, the most noteworthy examples are those that consider IR as process of uncertain inference. The concept of relevance is interpreted in a different way, where it can be extended and defined with respect, not only to a query formulation, but also to an information need.

The models of both categories are presented separately, but using a common formalism and, as much as possible, to the same level of detail.

2.3 Probabilistic relevance models

The main task of IR systems based upon relevance models is to evaluate a probability of a document being relevant. This is done by estimating the probability $P(R \mid q_k, d_i)$ for every document d_i in the collection, which is a difficult problem. The estimation problem can only be tackled by means of simplifying assumptions. Two kinds of approaches have been developed to deal with such assumptions: model-oriented and description-oriented.

Model-oriented approaches are based upon some probabilistic independence assumptions concerning the elements used in representing[3] the documents or the queries. The probabilities of these individual representation elements are estimated, and, by means of the independence assumptions, the probabilities of the document representations are estimated from them. The Binary Independence Indexing and Retrieval models (sections 2.3.3 and 2.3.2), and the n-Poisson model (Section 2.3.8) are examples of this approach.

Description-oriented approaches are more heuristic in nature. Given the representation of queries and documents, a set of features for query-document pairs is defined (e.g., occurrence frequency information), that allows each query-document pair in the collection to be mapped on to these features. Then, by means of some training data containing query-document pairs together with their corresponding relevance judgements, the probability of relevance is estimated with respect to these features. The best example of the application of this approach is the Darmstadt Indexing model (Section 2.3.4). However, a new model whose experimental results are not yet known, has been proposed by Cooper et al. [CGD92]. These models exploit the mapping between representations and descriptions that we introduced in Section 2.2.2.

2.3.1 Probabilistic Modelling as a decision strategy

The use of probabilities in IR was advanced in 1960 by Maron and Kuhns [MK60]. In 1976, Robertson and Sparck Jones went further by showing the powerful contribution of probability theory to model IR. The probabilistic model was theoretically finalised by van Rijsbergen in [vR79], chapter 6. The focus of the model is on its analysis as a decision strategy based upon a loss or risk function.

[3]Depending on the complexity of the models, the probabilities to be estimated can be with respect to the representations or the descriptions. But for clarity of expression, we will refer to the representations only, unless otherwise stated.

Referring to the conceptual model described in Section 2.2.2, it is assumed that the representation and the description methods for queries and documents are the same. Queries and documents are described by sets of index terms. Let $T = \{t_1, \ldots, t_n\}$ denote the set of terms used in the collection of documents. We represent the query q_k with terms belonging to T. Similarly, we represent a document d_j as the set of terms occurring in it. If we use a binary representation then d_j is represented as the binary vector $\vec{x} = (x_1, \ldots, x_n)$ with $x_i = 1$ if $t_i \in d_j$ and $x_i = 0$ otherwise. The query q_k is represented in the same manner.

The basic assumption, common to most models described in Section 2.3, is that the distribution of terms within the document collection provides information concerning the relevance of a document to a given query. This is because it is assumed that terms are distributed differently in relevant and non-relevant documents. This is known as the *cluster hypothesis* (see [vR79] pp. 45-47). If the term distribution was the same within the sets of relevant and non-relevant documents then it would not be possible to devise a discrimination criterion between them. In which case, a different representation of the document information content would be necessary.

The term distribution provides information about the "probability of relevance" of a document to a query. If we assume binary relevance judgements, then the term distribution provides information about $P(R \mid q_k, d_j)$.

The quantity $P(R \mid q_k, \vec{x})$, with $\vec{x}$ as a binary document representation, cannot be estimated directly. Instead, Bayes' theorem is applied [Pea88]:

$$P(R \mid q_k, \vec{x}) = \frac{P(R \mid q_k) \cdot P(\vec{x} \mid R, q_k)}{P(\vec{x} \mid q_k)}$$

To simplify notation, we omit the q_k on the understanding that evaluations are with respect to a given query q_k. The previous relation becomes:

$$P(R \mid \vec{x}) = \frac{P(R) \cdot P(\vec{x} \mid R)}{P(\vec{x})}$$

where $P(R)$ is the prior probability of relevance, $P(\vec{x} \mid R)$ is the probability of observing the description $\vec{x}$ conditioned upon relevance having been observed, and $P(\vec{x})$ is the probability that $\vec{x}$ is observed. The latter is determined as the joint probability distribution of the n terms within the collection. The

$C_j(R, dec)$	retrieved	not retrieved
relevant document	0	λ_1
non relevant document	λ_2	0

Table 2.1: The cost of retrieving and not retrieving a relevant and non relevant document

above formula evaluates the "posterior" probability of relevance conditioned upon the information provided in the vector $\vec{x}$.

The provision of a ranking of documents by the PRP can be extended to provide an "optimal threshold" value. This can be used to set a cut-off point in the ranking to distinguish between those documents that are worth retrieving and those that are not. This threshold is determined by means of a *decision strategy*, whose associated *cost function* $C_j(R, dec)$ for each document d_j is described in Table 2.1.

The decision strategy can be described simply as one that minimises the average cost resulting from any decision. This strategy is equivalent to minimising the following *risk function*:

$$\mathcal{R}(R, dec) = \sum_{d_j \in D} C_j(R, dec) \cdot P(d_j \mid R)$$

It can be shown (see [vR79], pp. 115-117) that the minimisation of that function brings about an optimal partitioning of the document collection. This is achieved by retrieving only those documents for which the following relation holds:

$$\frac{P(d_j \mid R)}{P(d_j \mid \overline{R})} > \lambda$$

where

$$\lambda = \frac{\lambda_2 \cdot P(\overline{R})}{\lambda_1 \cdot P(R)}$$

2.3.2 The Binary Independence Retrieval model

In the previous section, it remains necessary to estimate the joint probabilities $P(d_j \mid R)$ and $P(d_j \mid \overline{R})$, that is $P(\vec{x} \mid R)$ and $P(\vec{x} \mid \overline{R})$ if we consider the binary vector document representation $\vec{x}$.

In order to simplify the estimation process, the components of the vector $\vec{x}$ are assumed to be stochastically independent when conditionally dependent upon R or $\overline{R}$. That is, the joint probability distribution of the terms in the document d_j is given by the following product of marginal probability distributions:

$$P(d_j \mid R) = P(\vec{x} \mid R) = \prod_{i=1}^{n} P(x_i \mid R)$$

and

$$P(d_j \mid \overline{R}) = P(\vec{x} \mid \overline{R}) = \prod_{i=1}^{n} P(x_i \mid \overline{R})$$

This *binary independence assumption*, is the basis of a model first proposed by Robertson and Spark Jones in 1976 [RS76]: the *Binary Independence Retrieval model* (BIR). The assumption has always been recognised as unrealistic.

Nevertheless, as pointed out by Cooper (Section 2.5), the assumption that actually underpins the BIR model is not that of binary independence, but that of the weaker assumption of *linked dependence*:

$$\frac{P(\vec{x} \mid R)}{P(\vec{x} \mid \overline{R})} = \prod_{i=1}^{n} \frac{P(x_i \mid R)}{P(x_i \mid \overline{R})}$$

This states that the ratio between the probabilities of $\vec{x}$ occurring in relevant and non relevant documents is equal to the product of the corresponding ratios of the single terms.

Considering the decision strategy of the previous section, it is now possible to obtain a decision strategy by using a logarithmic transformation to obtain a linear decision function:

$$g(d_j) = \log \frac{P(d_j \mid R)}{P(d_j \mid \overline{R})} > \log \lambda$$

To simplify notation, we define the following quantities:

$$p_j = P(x_j = 1 \mid R)$$

and

$$q_j = P(x_j = 1 \mid \overline{R})$$

which represent the probability of the jth term appearing in a relevant, and in a non relevant document, respectively. Clearly: $1 - p_j = P(x_j = 0 \mid R)$, and $1 - q_j = P(x_j = 0 \mid \overline{R})$. This gives:

$$P(\vec{x} \mid R) = \prod_{j=1}^{n} p_j^{x_j} \cdot (1 - p_j)^{1-x_j}$$

and

$$P(\vec{x} \mid \overline{R}) = \prod_{j=1}^{n} q_j^{x_j} \cdot (1 - q_j)^{1-x_j}$$

Substituting the above, gives:

$$\begin{aligned} g(d_i) &= \textstyle\sum_{j=1}^{n} (x_j \cdot \log \frac{p_j}{q_j} + (1 - x_j) \cdot \log \frac{1-p_j}{1-q_j}) \\ &= \textstyle\sum_{j=1}^{n} c_j \; x_j + C \end{aligned}$$

where:

$$c_j = \log \frac{p_j \cdot (1 - q_j)}{q_j \cdot (1 - p_j)}$$

and

$$C = \sum_{j=1}^{n} \log \frac{1 - p_j}{1 - q_j}$$

This formula gives the RSV of document d_j for the query under consideration. Documents are ranked according to their RSV and presented to the user. The cut-off value λ can be used to determine the point at which the display of the documents is stopped, although, the RSV is generally used only to rank the entire collection of documents. In a real IR system, the presentation of documents ordered on their estimated probability of relevance to a query matters more than the actual value of those probabilities. Therefore, since the value of C is constant for a specific query, we need only consider the value of c_j. This value, or more often the value $\exp(c_j)$, is called the *term relevance weight* (TRW), and indicates the term's capability to discriminate relevant from non relevant documents. As it can be seen, in the BIR model term relevance weights contribute "independently" to the relevance of a document.

To apply the BIR model, it is necessary to estimate the parameters p_j and q_j for each term used in the query. This is performed in various ways, depending upon the amount of information available. The estimation can be retrospective or predictive. The first is used on test collections where the relevance assessments are known. The second is used with normal collection where parameters are estimated by means of relevance feedback from the user.

There is another technique, proposed by Croft and Harper [CH79], that uses a collection information to make estimates and does not use relevance information. Let us assume that the IR system has already retrieved some documents for the query q_k. The user is asked to give relevance assessments for those documents, from which the parameters of the BIR are estimated. If we also assume to be working in the retrospective case, then we know the relevance value of all individual documents in the collection. Let a collection have N documents, R of which are relevant to the query. Let n_j denote the number of documents in which the term x_j appears, amongst which, only r_j are relevant to the query. The parameters p_j and q_j can then be estimated as follows:

$$\hat{p_j} = \frac{r_j}{R}$$

and

$$\hat{q_j} = \frac{n_j - r_j}{N - R}$$

These give:

$$TRW_j = \frac{\frac{r_j}{R-r_j}}{\frac{n_j-r_j}{N-n_j-R+r_j}}$$

This approach is possible only if we have relevance assessments for all documents in the collection, i.e. where we know R and r_j. According to Croft and Harper, given that the only information concerning the relevance of documents is that provided by a user through relevance feedback, predictive estimations should be used. Let $\tilde{R}$ denote the number of documents judged relevant by the user. Further, let $\tilde{r_j}$ be the number of those documents in which the term x_j occurs. We can then combine this with the estimation technique of [Cox70].

$$T\tilde{RW}_j = \frac{\frac{\tilde{r_j}+0.5}{\tilde{R}-\tilde{r_j}+0.5}}{\frac{n_j-\tilde{r_j}+0.5}{N-n_j-\tilde{R}+\tilde{r_j}+0.5}}$$

Usually, the relevance information given by a user is limited and is not sufficiently representative of the entire collection. Consequently, the resulting estimates tend to lack precision. As a partial solution, one generally simplifies by assuming p_j to be constant for all the terms in the indexing vocabulary. The value $p_j = 0.5$ is often used, which gives a TRW that can be evaluated easily:

$$T\tilde{\tilde{RW}}_j = \frac{N - n_j}{n_j}$$

For large N, i.e. large collections of documents, this expression can be approximated by the "inverse document frequency" $IDF_j = \log\ N/n_j$. This is widely used in IR to provide an intuitive discrimination power of a term in a document collection.

2.3.3 The Binary Independence Indexing model

The *Binary Independence Indexing model* (BII model) is a variant of the BIR model. Where the BIR model regards a single query with respect to the

entire document collection, the BII model regards one document in relation to a number of queries. The indexing weight of a term is evaluated as an estimate of the probability of relevance of that document with respect to queries using that term. This idea was first proposed in Maron and Kuhns's indexing model [MK60].

In the BII, the focus is on the query representation, which we assume to be a binary vector $\vec{z}$. The dimension of the vector is given by the set of all terms T which could be used in a query, and $z_j = 1$ if the term represented by that element is present in the query, $z_j = 0$ otherwise[4]. In this model, the terms weights are defined in terms of frequency information derived from queries; that is, an explicit document representation is not required. We will only assume that there is a subset of terms that can be used to represent any document, and that will be given weights with respect to a particular document.

The BII model seeks an estimate of the probability $P(R \mid \vec{z}, d_j)$ that the document d_j will be judged relevant to the query represented by $\vec{z}$. To use the same formalism as the previous section, we use $\vec{x}$ to denote the document representation. So far this model looks very similar to the BIR; the difference lies with the application of Bayes' theorem as follows:

$$P(R \mid \vec{z}, \vec{x}) = \frac{P(R \mid \vec{x}) \cdot P(\vec{z} \mid R, \vec{x})}{P(\vec{z} \mid \vec{x})}$$

$P(R \mid \vec{x})$ is the probability that the document represented by $\vec{x}$ will be judged relevant to an arbitrary query. $P(\vec{z} \mid R, \vec{x})$ is the probability that the document will be relevant to a query with representation $\vec{z}$. As $\vec{z}$ and $\vec{x}$ are assumed to be mutually independent, $P(\vec{z} \mid \vec{x})$ reduces to the probability that the query $\vec{z}$ will be submitted to the system $P(\vec{z})$.

To proceed from here, some simplifying assumptions must be made:

1. The conditional distribution of terms in all queries is independent. This is the classic "binary independence assumption", from which the model's name arises:

$$P(\vec{z} \mid R, \vec{x}) = \prod_{i=1}^{n} P(z_i \mid R, \vec{x})$$

[4]As a consequence, two different information needs (i.e., two queries) using the same set of terms will produce the same ranking of documents.

2. The relevance of a document with representation $\vec{x}$ with respect to a query $\vec{z}$ depends only upon the terms used by the query (i.e., those with $z_i = 1$) and not upon other terms.

3. With respect to a specific document, for each term not used in the document representation, we assume:

$$P(R \mid z_i, \vec{x}) = P(R \mid \vec{x})$$

Now, applying the first assumption to $P(R \mid \vec{z}, \vec{x})$, we get:

$$P(R \mid \vec{z}, \vec{x}) = \frac{P(R \mid \vec{x})}{P(\vec{z} \mid \vec{x})} \cdot \prod_{i=1}^{n} P(z_i \mid R, \vec{x})$$

by applying the second assumption and Bayes' theorem, we get the ranking formula:

$$\begin{aligned} P(R \mid \vec{z}, \vec{x}) &= \frac{\prod_i P(z_i)}{P(\vec{z})} \cdot \prod_{i=1}^{n} \frac{P(R|z_i,\vec{x})}{P(R|\vec{x})} \\ &= \frac{\prod_i P(z_i)}{P(\vec{z})} \cdot P(R \mid \vec{x}) \cdot \prod_{z_i=1} \frac{P(R|z_i=1,\vec{x})}{P(R|\vec{x})} \cdot \prod_{z_i=0} \frac{P(R|z_i=0,\vec{x})}{P(R|\vec{x})} \end{aligned}$$

The value of the first fraction is a constant c for a given query, so there is no need to estimate it for ranking purposes. In addition, by applying the third assumption, the third fraction becomes equal to 1, and we obtain:

$$P(R \mid \vec{z}, \vec{x}) = c \cdot P(R \mid \vec{x}) \cdot \prod_{t_i \in \vec{z} \cap \vec{x}} \frac{P(R \mid t_i, \vec{x})}{P(R \mid \vec{x})}$$

There are a few problems with this model. The use of the third assumption is in contrast with experimental results reported by Turtle [Tur90], who demonstrates the advantage of assigning weights to query terms not occurring in a document. Moreover, the second assumption is called into question by Robertson et al. [RS76]. They proved experimentally the superiority of a ranking approach in which the probability of relevance is based upon both the presence and the absence of query terms in documents. The results suggest that the BII model might obtain better results if it were, for example, used together with a thesaurus or a set of term-term relations. This would enable the use of document terms not present in the query, but related in some way to those that were.

Fuhr [Fuh92b] pointed out that, in its present form, the BII model is hardly an appropriate model because, in general, there is not enough relevance information available to estimate the probability $P(R \mid t_i, \vec{x})$ for specific term-document pairs. To partially overcome this problem, one can assume that a document consists of independent components to which the indexing weights relate. However, experimental evaluations of this strategy have shown only average retrieval results [Kwo90].

Robertson et al. proposed a model that provides a unification of the BII and BIR models [RMC82]. The proposed model, simply called Model 3 (as opposed to the BII model called Model 1 and the BIR model called Model 2), enables us to combine the two retrieval strategies of the BII and the BIR models, thus providing a new definition of probability of relevance that unifies those of the BII and BIR models. In the BII model the probability of relevance of a document given a query is computed relative to evidence consisting of the properties of the queries for which that document was considered relevant, while in the BIR model it is computed relative to the evidence consisting of the properties of documents considered relevant by that same query. Model 3 enables us to use both forms of evidence. Unfortunately, a computationally treatable estimation theory fully faithful to Model 3 has not been proposed. The Model 3 idea has been explored later by Fuhr [Fuh89] and Wong and Yao [WY89] (see Section 2.3.5).

2.3.4 The Darmstadt Indexing model

The basic idea of the *Darmstadt Indexing approach* (DIA) is to use long-term learning of indexing weights from users' relevance judgements [FK84, BFK+88, FB91]. It can be seen as an attempt to develop index term specific estimates based upon the use of index terms in the learning sample.

DIA attempts to estimate $P(R \mid x_i, q_k)$ from a sample of relevance judgements of query-document or term-document pairs. This approach, when used for indexing, associates a set of heuristically selected attributes to each term-document pair, rather than estimating the probability associated with an index term directly (examples are given below). The use of an attribute set reduces the amount of training data required and allows the learning to be collection specific. However, the degree to which the resulting estimates are term specific depends critically upon the particular attributes used.

The indexing performed by the DIA is divided in two steps: a description step and a decision step.

In the *description step* relevance descriptions for term-document pairs $(x_i, \vec{x})$ are formed. These relevance descriptions $s(x_i, \vec{x})$ [5], comprise a set of attributes considered important for the task of assigning weights to terms with respect to documents. A relevance description $s(x_i, \vec{x})$ contains values of attributes of the term x_i, of the document (represented by $\vec{x}$) and of their relationships. This approach does not make any assumptions about the structure of the function s or about the choice of attributes. Some examples of attributes which could be used by the description function are:

- frequency of occurrence of term x_i in the document,
- inverse frequency of term x_i in the collection,
- information about the location of the occurrence of term x_i in the document, or
- parameters describing the document, e.g. its length, the number of different terms occurring in it, etc.

In the *decision step*, a probabilistic index weight based on the previous data is assigned. This means that we estimate $P(R \mid s(x_i, \vec{x}))$ and not $P(R \mid x_i, \vec{x})$. In the latter case, we would have regarded a single document d_j (or $\vec{x}$) with respect to all queries containing x_i, as in the BII model. Here, we regard the set of all query-document pairs in which the same relevance description s occurs. The interpretation of $P(R \mid s(x_i, \vec{x}))$ is therefore that of the probability of a document being judged relevant to an arbitrary query, given that a term common to both document and query has a relevance description $s(x_i, \vec{x})$.

The estimates of $P(R \mid s(x_i, \vec{x}))$ are derived from a learning sample of term-document pairs with attached relevance judgements derived from the query-document pairs. If we call this new domain L, we have:

$$L \subset \underline{D} \times \underline{Q} \times \Re$$

or

$$L = \{(q_k, d_j, r_{kj})\}$$

[5] These are similar to those used in pattern recognition.

By forming relevance descriptions for the terms common to queries and documents for every query-document pair in L, we get a multi-set of relevance descriptions with relevance judgements:

$$L^x = [(s(x_i, d_j), r_{kj}) \mid x_i \in q_k \cap d_j \wedge (q_k, d_j, r_{kj}) \in L]$$

Using this set, it would be possible to estimate $P(R \mid s(x_i, \vec{x}))$ as the relative frequency of those elements of L^x with the same relevance description. Nevertheless, the technique used in DIA makes use of an *indexing function*, because it provides better estimates through the use of additional plausible assumptions about the indexing function. In [FB91], various linear indexing functions estimated by least squares polynomial were used, while in [FB93] a logistic indexing function estimated by maximum likelihood was attempted. Experiments were performed using both a controlled and a free term vocabulary.

The experimental results on the standard test collections indicate that the DIA approach is often superior to other indexing methods. The more recent, but only partial, results obtained using the TREC collection [FB93] tend to support this conclusion.

2.3.5 The Retrieval with Probabilistic Indexing model

The *Retrieval with Probabilistic Indexing* (RPI) model described in [Fuh89] takes a different approach from other probabilistic models. This model assumes that we use not only a weighting of index terms with respect to the document but also a weighting of query terms with respect to the query. If we denote w_{mi} the weight of index term x_i with respect to the document $\vec{x}_m$, and v_{ki} the weight of query term $z_i = x_i$ with regard to the query $\vec{z}_k$, then we can evaluate the following scaler product and use it as retrieval function:

$$r(\vec{x}_m, \vec{z}_k) = \sum_{\{x_m = z_k\}} w_{mi} \cdot v_{ki}$$

Wong and Yao [WY89] give an utility theoretic interpretation of this formula for probabilistic indexing. Assuming we have a weighting of terms with respect to documents (similar to those, for example, of BII or DIA), the weight v_{ki} can be regarded as the utility of the term t_i, and the retrieval function $r(d_m, q_k)$ as the expected utility of the document with respect to the

query. Therefore, $r(d_m, q_k)$ does not estimate the probability of relevance, but it has the same utility theoretic justification as the PRP.

RPI was developed especially for combining probabilistic indexing weighting with query term weighting based, for example, on relevance feedback. As a result, its main advantage is that it is suitable for application to different probabilistic indexing schemes.

2.3.6 The Probabilistic Inference model

Wong and Yao in [WY95] extend the work reported in [WY89] by using an epistemological view of probability, from where they proposed a probabilistic inference model for IR. With the epistemic view of probability theory, the probabilities under consideration are defined based on semantic relationships between documents and queries. The probabilities are interpreted as degrees of beliefs.

The general idea of the model starts with the definition of a concept space, which can be interpreted as the knowledge space in which documents, index terms, and user queries are represented as propositions. For example: the proposition d is the knowledge contained in the document; the proposition q is the information need requested; and the proposition $d \cap q$ is the portion of knowledge common to d and q.

An epistemic probability function P is defined on the concept space. For example, $P(d)$ is the degree to which the concept space is covered by the knowledge contained in the document and $P(d \cap q)$ is the degree to which the concept space is covered by the knowledge common to the document and the query.

Based on these probabilities, different measures can be constructed to evaluate the relevance of documents to queries, offering different interpretations of relevance, thus leading to different approaches to model IR. I discuss two of them. The first one is:

$$\Psi(d \rightarrow q) = P(q|d) = \frac{P(d \cap q)}{P(d)}$$

$\Psi(d \rightarrow q)$ can be considered as a measure of precision of the document with respect to the query, and is defined as the probability that a retrieved document is relevant. A precision-oriented interpretation of relevance should

be used, for example, when a user is interested in locating a specific piece of information. A second measure is:

$$\Psi(q \rightarrow d) = P(d|q) = \frac{P(q \cap d)}{P(q)}$$

$\Psi(q \rightarrow d)$ is considered as a recall index of the document with respect to the query, and is defined as the probability that a relevant document is retrieved. A recall-oriented measure should be used when the user is writing a review paper on a particular subject, and is interested in finding as many papers as possible on the subject.

Depending of the relationships between concepts, different formulations of $\Psi(d \rightarrow q)$ and $\Psi(q \rightarrow d)$ are obtained. For example, suppose that the concept space is $t_1 \cup \ldots \cup t_n$ where the basic concepts are (pairwise) disjoint; i.e., $t_i \cap t_j = \emptyset$ for $i \neq j$. It can be proven that, taking a $t \in \{t_1, \ldots t_n\}$:

$$\Psi(d \rightarrow q) = \frac{\sum_t P(d \cap q|t)P(t)}{P(d)}$$

$$\Psi(q \rightarrow d) = \frac{\sum_t P(d \cap q|t)P(t)}{P(q)}$$

Wong and Yao work aims to provide a probabilistic evaluation of uncertain implications which have been advanced as a way to measure the relevance of documents to queries (see Section 2.4.1). Although measuring uncertain implications by a probability function is more restrictive than for example using the possible world analysis, the model proposed by Wong and Yao is both expressive and sound. For example, they show that the Boolean, fuzzy set, Vector Space and probabilistic models are special cases of their model. I will not go into the detail of this demonstration, but I refer to the cited articles.

2.3.7 The Staged Logistic Regression model

Cooper's *Staged Logistic Regression model* (SLR), proposed in [CGD92], is an attempt to overcome some problems present in the use of standard regression methods to estimate probabilities of relevance in IR. Cooper criticises

Fuhr's approaches, especially the DIA which require strong simplifying assumptions. He thinks (I include a longer explanation of his point of view in Section 2.5) that these assumptions inevitably distort the final estimate of the probability of relevance. He advocates a "model-free" approach to estimation. In addition, a more serious problem lies in the use of standard polynomial regression methods. Standard regression theory is based on the assumption that the sample values taken for the dependent variable are from a continuum of possible magnitudes. In IR, the dependent variable is usually dichotomous: a document is either relevant or non relevant. So standard regression is clearly inappropriate in such cases.

A more appropriate tool, according to Cooper, is *logistic regression*, a statistical method specifically developed for using dichotomous (or discrete) dependent variables. Related techniques were used with some success by other researchers, for example, Fuhr employed it in [FP91] and more recently in [FB93].

The method proposed by Cooper is based on the guiding notion of treating composite clues on at least two levels, an intra-clue level at which a predictive statistic is estimated separately for each composite clue[6], and an inter-clue level in which these separate statistics are combined to obtain an estimate of the probability of relevance for a query-document pair. As this proceeds in stages, the method is called Staged Logistic Regression (SLR). A two stage SLR would be as follows:

1. A statistical simplifying assumption is used to break down the complex joint probabilistic distribution of the composite clues. This assumption is called *linked dependence*. For example, assuming that we have only two clues, a positive real number K exists such that the following conditions hold true:

$$P(a, b \mid R) = K\ P(a \mid R) \cdot P(b \mid R)$$

$$P(a, b \mid \neg R) = K\ P(a \mid \neg R) \cdot P(b \mid \neg R)$$

It follows that:

$$\frac{P(a, b \mid R)}{P(a, b \mid \neg R)} = \frac{P(a \mid R)}{P(a \mid \neg R)} \cdot \frac{P(b \mid R)}{P(b \mid \neg R)}$$

[6] A simple clue could be, for example, the presence of an index term in a document. Clues need to be machine-detectable.

Generalising this result to the case of n clues and taking the "log odds" we obtain:

$$LogO(R \mid a_1, \ldots, a_n) = \\ = LogO(R) + \sum_{i=1}^{n} (LogO(R \mid a_i) - LogO(R))$$

This is used at retrieval time to evaluate the log odds of relevance for each document in the collection with respect to the query.

2. A logistic regression analysis on a learning sample is used to obtain an estimate of the terms on the right hand side of the previous equation. Unfortunately, the required learning sample is often only available within the environment of test collections, although it could be possible to use the results of previous good queries for this purpose.

   The estimation of $LogO(R)$ is quite straightforward using simple proportions. A more complex matter is the estimation of $LogO(R \mid a_i)$, when there are too few query-document pairs in the learning set with the clue a_i to yield estimates of $P(R \mid a_i)$ and $P(\neg R \mid a_i)$. To go beyond simple averaging, Cooper uses multiple logistic regression analysis. If we assume that the clue a_i is a composite clue, whose elementary attributes are $h_i, \ldots, h_m$ then we can estimate $LogO(R \mid a_i)$ as follows:

$$\begin{aligned} LogO(R \mid a_i) &= LogO(R \mid h_i, \ldots, h_m) \\ &= c_0 + c_1\ h_1, + \ldots + c_m\ h_m \end{aligned}$$

   To demonstrate how the logistic function comes into the model, the probability of relevance of a document can be expressed as:

$$P(R \mid h_i, \ldots, h_m) = \frac{e^{c_0 + c_1\ h_1, + \ldots + c_m\ h_m}}{1 + e^{c_0 + c_1\ h_1, + \ldots + c_m\ h_m}}$$

   Taking the log odds of both sides conveniently reduces this formula to the previous one.

3. A second logistic regression analysis, based on the same learning sample, is used to obtain another predictive rule for combining the composite clues and for correcting biases introduced by the simplifying assumption.

The linked dependence assumption tends to inflate the estimates for documents near the top of the output ranking whenever the clues on which the estimates are based are strongly interdependent. To help correct this, a second level logistic regression analysis is performed on the results of the first. It has the following form:

$$LogO(R \mid a_1, \ldots, a_n) = d_0 + d_1 \; Z + d_2 \; n$$

where $Z = \sum_{i=1}^{n}(LogO(R \mid a_i) - LogO(R))$ and n is the number of composite clues. More elaborate correcting equation might also be considered.

When a query is submitted to the system and a document is compared against it the technique in part 2 is applied to evaluate the log odds necessary to obtain Z. That is then employed in part 3 to adjust estimate of the log odds of relevance for the document.

This approach seems flexible enough to handle almost any type of probabilistic retrieval clues likely to be of interest, and is especially appropriate when the retrieval clues are grouped or composite. However, the effectiveness of the methodology remains to be determined empirically, and its performance compared with other retrieval methods. An experimental investigation is currently under way by Cooper, and the use of logistic regression has also been investigated by Fuhr, as reported in the proceedings of the TREC-1 Conference [FB93].

2.3.8 The N-Poisson indexing model

This probabilistic indexing model is an extension to n-dimensions of the 2-*Poisson model* proposed by Bookstein et al. in 1974 [BS74]. In its *2-dimensional* form the model is based upon the following assumption. If the number of occurrences of a term within a document is different depending upon whether the document is relevant or not, and if the number of occurrences of that term can be modelled using a known distribution, then it is possible to decide if a term should be assigned to a document by determining which of the two distributions the term belongs to. The 2-Poisson model resulted from a search for the statistical distribution of occurrence of potential index terms in a collection of documents.

We can extend the above idea to the *n-dimensional* case. We suppose there are n classes of documents in which the term x_i appears with different frequencies according to the extents of coverage of the topic related to that specific term. The distribution of the term within each class is governed by a single Poisson. Given a term x_i, and a document class for that term K_{ij}, and the expectation of the number of occurrences of that term in that class λ_{ij}, then the probability that a document contains l occurrence of x_i, i.e. that $tf(x_i) = l$, given that it belongs to the class K_{ij}, is given by:

$$P(tf(x_i) = l \mid \vec{x} \in K_{ij}) = \frac{\lambda_{ij}^l}{l!} e^{-\lambda_{ij}}$$

Extending this result, the distribution of a certain term within the whole collection of documents is governed by a sum of Poisson distributions, one for each class of coverage. In other words, if we take a document at random in the collection, whose probability of belonging to class K_{ij} is p_{ij} then the probability of having l occurrences of term x_i is:

$$P(tf(x_i) = l) = \sum_{j=1}^{n} p_{ij} \, e^{-\lambda_{ij}} \frac{\lambda_{ij}^l}{l!}$$

This result can be used with a Bayesian inversion to evaluate $P(\vec{x} \in K_{ij} \mid tf(x_i = l))$ for retrieval purposes. The parameters λ_{ij} and p_{ij} can be estimated without feedback information by applying statistical techniques to the document collection.

Experiments have shown that the performance of this model is not always consistent. Some experiments performed by Harter [Har75] on a 2-Poisson showed that a significant number of "good" index terms were 2-Poisson, but they did not provide conclusive evidence of the validity of the n-Poisson model. These results were co-validated by Robertson et al. [RW94]. They demonstrated considerable performance improvements by using some effective approximations to the 2-Poisson model on the TREC collection. Other research investigated the possibility of using a 3-Poisson, and lastly Margulis [Mar92, Mar93] investigated the generalised n-Poisson model on several large full text document collections. His findings were more encouraging than those of the previous work. He determined that over 70% of frequently occurring words were indeed distributed according to a n-Poisson distribution. Further, he found that the distribution of most n-Poisson words had relatively

few single Poisson components, instead, usually two, three, or four. He concluded suggesting that his study provides strong evidence that the n-Poisson distribution could be used as a basis for accurate statistical modelling of large document collections. However, to date, the n-Poisson approach lacks work on retrieval strategies based upon the results gained so far.

2.4 Uncertain inference models

The models presented in this section are based on the idea that IR is a process of uncertain inference. Uncertain inference models are based on more complex forms of relevance that those used in relevance models, which are based mainly upon statistical estimations of the probability of relevance. With uncertain inference models, information not present in the query formulation may be included in the evaluation of the relevance of a document. Such information might be domain knowledge, knowledge about the user, user's relevance feedback, etc. The estimation of the probabilities $P(R \mid q_k, d_i, K)$ involves the representation of the knowledge K.

Another characteristic of uncertain inference models is that they are not as strongly collection-dependent as relevance models. Parameters in relevance models are only valid for the current collection, while inference models can use knowledge of the user or the application domain that can be useful with many other collections.

This research area is promising in that it is attempting to move away from the traditional approaches, and may provide the breakthrough that appears necessary to overcome the limitations of current IR systems.

There are two main types of uncertain inference models. The first is based on non-classical logic, to which probabilities are mapped (Section 2.4.1), and the second is based on Bayesian inferences (Section 2.4.2).

2.4.1 A non-classical logic for IR

In 1986, Van Rijsbergen proposed a paradigm for probabilistic IR in which IR was regarded as a process of uncertain inference [vR86]. The paradigm is based on the assumption that queries and documents can be regarded as logical formulae, and to answer a query, an IR system must prove the query from the documents. This means that a document is relevant to a query only

if it implies the query; in other words, if the logical formula $d \rightarrow q$ can be proven to *hold*. The proof may use additional knowledge K; in that case, the logical formula is then rewritten as $(d, K) \rightarrow q$.

The introduction of uncertainty comes from the consideration that a collection of documents cannot be considered as a consistent and a complete set of statements. In fact, documents in the collection could contradict each other in any particular logic, and not all the necessary knowledge is available. It has been shown [vR86, Lal97] that classical logic, the most commonly used logic, is not adequate to represent query and documents because of the intrinsic uncertainty present in IR[7]. Therefore, Van Rijsbergen proposes the *logical uncertainty principle* [vR86]:

> "Given any two sentences x and y; a measure of the uncertainty of $y \rightarrow x$ related to a given data set is determined by the minimal extent to which we have to add information to the data set, to establish the truth of $y \rightarrow x$"

The principle says nothing about how "uncertainty" and "minimal" might be quantified. However, in his paper, Van Rijsbergen suggested an information-theoretic approach. This idea has been followed by Nie et al. [NLB95] and Lalmas [vRL96]. However, that work is somewhat beyond the scope of this chapter.

More close to this survey, Van Rijsbergen [vR89] later proposed to estimate $P(d \rightarrow q)$ by *imaging*. Imaging formulates probabilities based on a "possible worlds" semantics [Sta81]. According to this semantics, a document is represented by a possible world w; i.e. a set of propositions with associated truth values. Let τ denote a logical truth function, then $\tau(w, y)$ denotes the truth of the proposition y in the world w. Further, let $\sigma(w, y)$ denote the world most similar to w where y is true. Then, $y \rightarrow x$ is true at w if and only if x is true at $\sigma(w, y)$.

Imaging uses this notion of most similar worlds to estimate $P(y \rightarrow x)$. Every possible world w has a probability $P(w)$, and the sum over all possible worlds is 1. $P(y \rightarrow x)$ is computed in the following way:

$$\begin{aligned} P(y \rightarrow x) &= \textstyle\sum_w P(w)\ \tau(w,\ y \rightarrow x) \\ &= \textstyle\sum_w P(w)\ \tau(\sigma(w, y),\ y \rightarrow x) \\ &= \textstyle\sum_w P(w)\ \tau(\sigma(w, y),\ x) \end{aligned}$$

[7]There are other reasons why classical logic is not adequate, but these are not relevant to this chapter (but see [Lal97]).

It remains undetermined how to evaluate the function σ on document representations, and further, how to assign a probability P to them.

In the above framework, the concept of relevance, does not feature. In [vR92] Van Rijsbergen proposed to evaluate the probability of relevance $P(R \mid q_k, d_i)$ using *Jeffrey's conditionalisation*. This conditionalisation, described as "Neo-Bayesianism" by Pearl [Pea90], allows conditioning to be based on evidence derived from the "passage of experience", where the evidence can be non-propositional in nature. A comprehensive treatise of Jeffrey's studies on probability kinematics, i.e. on how to revise a probability measure in the light of uncertain evidence or observation, can be found in [Jef65]. By means of the famous example of inspecting the colour of a piece of cloth by candlelight in that book, Van Rijsbergen introduced a form of conditioning that has many advantages over Bayesian conditioning. In particular, it enables conditioning on uncertain evidence, and allows order-independent partial assertion of evidence. Such advantages, despite some strong assumptions, convinced van Rijsbergen that this particular form of conditionalisation is more appropriate for IR than Bayesian conditionalisation. However, despite the appeal of Jeffrey's conditionalisation, the evaluation of the probability of relevance involves parameters, the estimation of which remain problematic.

In the same paper [vR92], Van Rijsbergen makes the connection between Jeffrey's conditionalisation and the Dempster-Shafer's Theory of Evidence [Dem68, Sha76]. This theory can be viewed as a generalisation of the Bayesian method (for example, it rejects the additivity rule), and have been used by some researchers to develop IR models (see [SH93, dSM93]).

2.4.2 The Inference Network model

When IR is regarded as a process of uncertain inference, then the calculation of the probability of relevance, and the general notion of relevance itself, becomes more complex. Relevance becomes related to the inferential process by which we find and evaluate a relation between a document and a query.

A probabilistic formalism for describing inference relations with uncertainty is provided by *Bayesian inference networks*, which have been described extensively in [Pea88] and [Nea90]. Turtle and Croft [Tur90, TC90, TC91] applied such networks to IR. Figure 2.2 depicts an example of such a network. Nodes represent IR entities such as documents, index terms, concepts, queries, and information needs. We can choose the number and kind of nodes we wish to

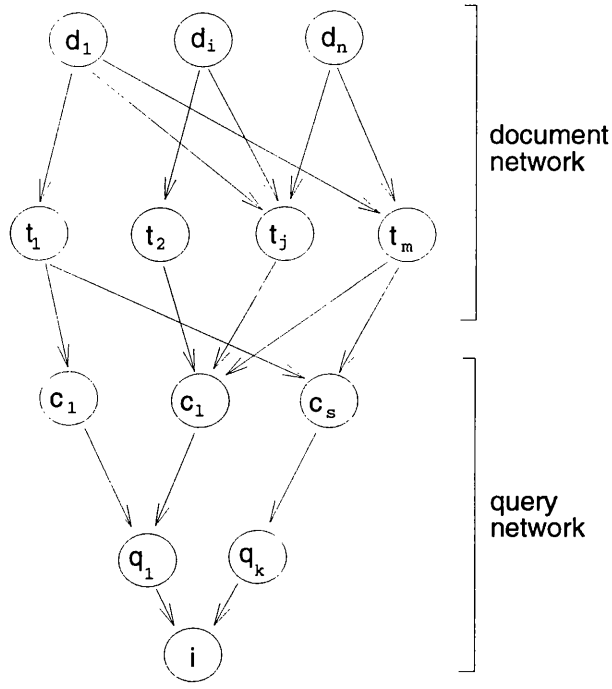

Figure 2.2: An inference network for IR.

use according to how complex we want the representation of the document collection or the information needs to be. Arcs represent probabilistic dependencies between entities. They represent conditional probabilities; that is, the probability of an entity being true given the probabilities of its parents being true.

The inference network is usually made up of two component networks: a document network and a query network. The document network represents the document collection. It is built once for a given collection and its structure does not change. A query network is built for each information need and can be modified and extended during each session by the user in a interactive and dynamic way. The query network is attached to the static document network in order to process a query.

In a Bayesian inference network, the truth value of a node depends only upon the truth values of its parents. To evaluate the strength of an inference chain going from one document to the query we set the document node d_i to "true" and evaluate $P(q_k = true \mid d_i = true)$. This gives us an estimate of $P(d_i \rightarrow q_k)$.

It is possible to implement various traditional IR models on this network by

introducing nodes representing Boolean operators or by setting appropriate conditional probability evaluation functions within nodes.

One particular characteristic of this model that warrants exploration is that multiple document and query representations can be used within the context of a particular document collection (e.g., a Boolean expression or a vector). Moreover, given a single information need, it is possible to combine results from multiple queries and from multiple search strategies.

The strength of this model comes from the fact that most classical retrieval models can be expressed in terms of a Bayesian inference network by estimating in different ways the weights in the inference network [TC92a]. Nevertheless, the characteristics of the Bayesian inference process itself, given that nodes (evidence) can only be binary (either present or not) limits its use to where "certain evidence" [Nea90] is available. The approach followed by van Rijsbergen (Section 2.4.1), which makes use of "uncertain evidence" by using Jeffrey's conditionalisation, therefore appears attractive.

2.5 Effective results from faulty models

Most of the probabilistic models presented in this chapter use simplifying assumptions to reduce the complexity related to the application of mathematical models to real situations. There are general risks inherent in the use of such assumptions. One such risk is that there may be *inconsistencies* between the assumptions laid down and the data to which they are applied.

Another is that there may be a *misidentification of the underlying assumptions*, i.e. the stated assumptions may not be the real assumptions upon which the derived model or resulting experiments are actually based. This risk was identified by Cooper [Coo95]. He identified the three most commonly adopted simplifying assumptions which are related to the statistical independence of documents, index terms, and information needs:

Absolute Independence

$$P(a, b) = P(a) \cdot P(b)$$

Conditional Independence

$$P(a, b \mid R) = P(a \mid R) \cdot P(b \mid R)$$

$$P(a, b \mid \neg R) = P(a \mid \neg R) \cdot P(b \mid \neg R)$$

These assumptions are interpreted differently whether a and b are regarded as properties of documents or of users.

Cooper pointed out how the combined use of the Absolute Independence assumption *and* either of the Conditional Independence assumptions yields logical inconsistencies. The combined use of these assumptions leads to the conclusion that $P(a, b, R) > P(a, b)$, which is contrary to the elementary laws of probability theory. Nevertheless, in most cases where these inconsistencies appeared, the faulty model used as the basis for experimental work has proved, on the whole, to be successful. Examples of this are given in [RS76] and [FB91].

The conclusion drawn by Cooper is that the experiments performed were actually based on somewhat different assumptions, which were, in fact, consistent. In some cases where the Absolute Independence assumption was used together with a Conditional Independence assumption, it seems that the required probability rankings could have been achieved on the basis of the Conditional Independence assumption alone. This is true of the model proposed by Maron et al. in [MK60]. In other cases, the Conditional Independence assumptions could be replaced by the single *linked dependence* assumption:

$$\frac{P(a, b \mid R)}{P(a, b \mid \neg R)} = \frac{P(a \mid R)}{P(a \mid \neg R)} \cdot \frac{P(b \mid R)}{P(b \mid \neg R)}$$

This is a considerably weaker assumption, and it is consistent with the Absolute Independence assumption. This is true of the SLR model presented in Section 2.3.7, and of the BIR model (whose name seems to lose appropriateness in the light of these results) presented in Section 2.3.2.

2.6 Further research

In the late nineties, we have come to realise that there is a leap to be made towards a new generation of IR systems; towards systems able to cope with increasingly demanding users, whose requirements and expectations continue to outstrip the progress being made in computing, storage, and transport technology. Faster machines and better interconnectivity enable access to

enormous amounts of information. This information is not only increasing in amount, but also in complexity; for example, structured hypertexts consisting of multiple media are becoming the norm. Until recently, research in Information Retrieval has been confined to the academic world. Things are changing slowly. The success of the TREC initiative (from [Har93] to [Har96]), particularly in terms of the interest shown by commercial organisations, demonstrates that there is a wider desire to produce sophisticated IR systems. The Web search engines, which have a high profile in the wider community, increasingly utilise probabilistic techniques. It can only be hoped that this increasing awareness and interest will stimulate new research.

The requirements of the next generation of IR systems include:

Multimedia documents The problem with multimedia document collections lies in the representation of the non-textual parts of documents, e.g. sounds, images, animations. Several approaches have been tried so far: they can be exemplified in the particular approach of attaching textual descriptions to non-textual parts, and the derivation of such descriptions by means of an inference process (e.g. [Dun91]). Nevertheless, such techniques avoid the real issue of directly handling the media. This applies not only to probabilistic models, but to all IR models.

Interactive retrieval Current IR systems, even those providing forms of relevance feedback for the user, are still based upon the traditional iterative batch retrieval approach. Even relevance feedback acts upon a previous retrieval run to improve the quality of the successive run [Har92c, Har92b]. We need real interactive systems, enabling a greater variety of interaction with the user than merely query formulation and relevance feedback [Cro87]. User profile information, analysis of browsing actions, or user modification of probabilistic weights, for example, could all be taken into consideration [CR87, CLC88, CLCW89, Tho89, Tho90a, Tho90b]. The subjective, contextual, and dynamic nature of relevance is now being recognised and incorporated into probabilistic models [CvR96].

Integrated text and fact retrieval There has been a steady development of the kinds of information being collected and stored in databases; notably, of text (unformatted data), and of 'facts' (formatted, often numerical, data). Demand is growing for the availability of systems capable of dealing with all types of data in a consistent and unified manner [Fuh92a, CST92, HW92, Fuh93].

Imprecise data The use of probabilistic modelling in IR is not only important for representing the document information content, but also for representing and dealing with vagueness and imprecision in the query formulation and with imprecision and errors in the textual documents themselves [Fuh90, TC92b]. For example, the increasing use of scanners and OCR in transferring documents from paper to electronic form, inevitably introduces imprecision (but see [SS88]).

2.7 Conclusions

The major concepts and a number of probabilistic IR models have been described. I am aware that new models are being developed as we speak. A survey is always a bit dated. However, I believe I have covered the most important and the most investigated probabilistic models of IR.

It is not easy to draw conclusions from a survey of thirty years of research. It is safe to conclude that good results have been achieved but more research is required since there is considerable room for improvement. Current generation probabilistic IR systems work quite well when compared with the Boolean systems that they are replacing. A novice user using natural language input with a current generation probabilistic IR system gets, on average, better performance than an expert user with a Boolean system on the same collection. Moreover, theoretically, the probabilistic approach to IR seems inherently suitable for the representation and processing of the uncertain and imprecise information that is typical of IR. I believe that, with development, it will be capable ultimately of providing an integrated, holistic, and theoretically consistent framework for the effective retrieval of complex information.

Part III

Theoretical Study

Chapter 3

Information Retrieval by Logical Imaging

The evaluation of an implication by Imaging is a logical technique developed in the framework of modal logic. Its interpretation in the context of a "possible worlds" semantics is very appealing for IR. In 1989, Van Rijsbergen suggested its use for solving one of the fundamental problems of logical models in IR: the evaluation of the implication $d \rightarrow q$ (where d and q are respectively a document and a query representation). Since then, others have tried to follow that suggestion proposing models and applications, though without much success. Most of these approaches had as their basic assumption the consideration that "a document is a possible world". I propose instead an approach based on a completely different assumption: "a term is a possible world". This approach enables the exploitation of term–term relationships which are estimated using an information theoretic measure.

3.1 The use of non-classical logic in Information Retrieval

The use of a probabilistic model in IR assures that we can obtain "optimal retrieval performance" once we rank documents according to their probability of relevance with regards to a query [Rob77]. However, this principle, called *The Probability Ranking Principle*, refers only to "optimal retrieval", which is different from "perfect retrieval". Optimal retrieval can be defined precisely for probabilistic IR because optimality can be proved theoretically,

owing to a provable relationship between ranking and the probabilistic interpretation of precision and recall [Rob76]. Perfect retrieval relates to the objects of the IR systems themselves, i.e. documents and information needs, but as IR systems use representations of these objects, perfect retrieval is not an appropriate goal for computer-based systems and cannot be achieved experimentally. Despite that and despite a few criticisms [Coo94], probabilistic models based on the Probability Ranking Principle have been shown to give the highest levels of retrieval effectiveness currently available [Fuh92b].

Although there are some operative IR systems based on probabilistic or semi-probabilistic models, there are still obstacles to getting probabilistic models accepted in the commercial IR world. One major obstacle is that of finding methods for estimating the probabilities of relevance that are both effective and computationally efficient. Past and present research has made much use of formal probability theory and statistics in order to solve the problems of estimation, see for example [CH79, FB91, WY89]. In mathematical terms the problem consists of estimating the probability $P(R \mid q, d)$, that is the probability of relevance given a query q and a document d. This estimate should be performed for every document in the collection, and documents should then be ranked according to this measure. This is a difficult task because of the large number of variables involved in the representation of documents in comparison with the small amount of feedback data available about the relevance of documents, a problem sometimes referred to as the "curse of dimensionality" [Eft96, vR79].

In 1986 Van Rijsbergen [vR86] proposed to use an estimation techniques based on the use of non-classical conditional logic. This enables the estimation of $P(R \mid q, d)$ by the evaluation of $P(d \rightarrow q)$, therefore using the probability of a conditional to estimate the conditional probability.

There are two main reasons behind the choice of $P(d \rightarrow q)$ to evaluate $P(R \mid q, d)$. The first one is that in this way we can separate the process of revising probabilities from the logic, the second is that we can separate the treatment of relevance from the treatment of documents and queries. In order to evaluate $P(R \mid q, d)$ we would need to resort to Bayes' Theorem:

$$P(R \mid q, d) \propto P(q \mid R, d)\ P(R)$$

Another way of putting this is that P is revised to a different probability function P_R in the light of information about relevance:

$$P(q \mid R, d) = P_R(q \mid d)$$

However, putting it in this way, it is clear that two users with differing ideas of relevance but submitting the same query can expect to get different probability of relevance for the same document, i.e. user one would get $P_R^1(q \mid d)$ and user two $P_R^2(q \mid d)$. This means that the probability of relevance can be revised in different ways. Moreover, what about the case of same relevance judgements but different queries? The probabilistic model does not deal with it directly, but the evaluation of $P(d \rightarrow q)$ enables one to address these problems.

According to Van Rijsbergen's view, the logical implication $d \rightarrow q$ is not one of material implication, the usual truth-functional connective $d \supset q$, which is always true in all cases except when d is true and q is false, but is based on a non-classical notion. The evaluation of the probability of the implication should be based on the following *logical uncertainty principle*:

> "Given any two sentences x and y; a measure of the uncertainty of $y \rightarrow x$ related to a given data set is determined by the minimal extent to which we have to add information to the data set, to establish the truth of $y \rightarrow x$."

The logical uncertainty principle initiated a new line followed by many researchers; see for example the work of Nie [Nie89], Chiaramella and Chevallet [CC92], Bruza [Bru93], and Huibers [HLvR96]. However, in the original 1986 paper Van Rijsbergen did not provide an indication about how "uncertainty" and "minimal" might be quantified. Only a few years later Van Rijsbergen proposed to estimate the probability of the conditional by a process called *Logical Imaging* (or simply Imaging), but without explicitly defining a technique that could be used operatively [vR89]. The Imaging technique was explored by Crestani and Van Rijsbergen and a model called *Retrieval by Logical Imaging* has now been defined in detail [CvR95].

I propose a technique called *Retrieval by Logical Imaging* (RbLI), that is based on the ideas suggested by Van Rijsbergen. It enables the evaluation of $P(d \rightarrow q)$ and $P(q \rightarrow d)$ by Imaging according to a possible worlds semantics where a term is considered as a possible world. This technique exploits term–term relationships in retrieval by means of an accessibility relation between worlds based on the Expected Mutual Information Measure (EMIM) estimated as described in [vR77].

The chapter is structured as follows. In Section 3.2 I give a brief explanation of what the Imaging process is all about, whilst Section 3.3 presents how

Imaging can be used in IR. Section 3.4 deals with the problems related to the implementation of RbLI. Section 3.5 reports on some experiments aiming at evaluating the effectiveness of the proposed technique. Related work is reported in Section 3.6, while the conclusions are described in Section 3.7.

3.2 Imaging and possible worlds semantics

Imaging is a process developed in the framework of Modal Logic. It enables the evaluation of a conditional sentence without explicitly defining the operator "$\rightarrow$". What it requires is a clustering on the space of events (worlds) by means of a primitive relation of neighbourhood. This semantics is called *possible worlds semantics* and was proposed by Kripke in [Kri71]. According to this semantics the truth value of the conditional $y \rightarrow x$ in a world w is equivalent to the truth value of the consequent x in the closest world w_y where the antecedent y is true. The identification of the closest world is done using the clustering. Ties at this stage, if they occur, are broken at random, to ensure uniqueness of the closest world (but see [Gö82] for a generalisation). The passage from a world to another world can be regarded as a form of belief revision, and the passage from a world to its closest is therefore equivalent to the least drastic revision of one's beliefs. Using this process it is possible to implement the logical uncertainty principle described in Section 3.1. Imaging can be extended to the case where we have a probability distribution on the worlds [Lew81]. A probability distribution over the worlds can be regarded as a measure of the prior uncertainty (or certainty) associated with the beliefs. In this case there is a shift of the original probability P of the world w to the closest world w_y where y is true. Probability is neither created nor destroyed, it is moved from a "not-y-world" to a "y-world" to derive a new probability distribution P'. This process is called *deriving P' from P by imaging on y.*

To explain in more detail how the Imaging process works, we need to use a little algebra and to introduce some terminology. The explanation will be in terms of the possible worlds semantics and it refers to the interpretation given by Stalnaker [Sta81].

Suppose we have a set of possible worlds W. Let w_y be the world most similar[1] (the closest if we have a distance metric) to w where y is true, then

[1]The notion of world and similarity used here is the standard one introduced by David Lewis in [Lew86].

$y \rightarrow x$ will be true at w if and only if x is true at w_y. Now let:

$$w(y) = \begin{cases} 1 & \text{if } y \text{ is true at } w \\ 0 & \text{otherwise} \end{cases}$$

then we have:

$$w(y \rightarrow x) = w_y(x)$$

where:

$$w_y(x) = \begin{cases} 1 & \text{if } x \text{ is true at } w_y \\ 0 & \text{otherwise} \end{cases}$$

Now we assume a probability distribution over the set of possible worlds W so that, according to the classical rules of probability, we have:

$$\sum_W P(w) = 1$$

Hence we define $P(y)$ as follows:

$$P(y) = \sum_W P(w)\ w(y)$$

From this probability distribution we can derive a new probability distribution P' so that:

$$P'(w') = \sum_W P(w)\ I(w', w)$$

where:

$$I(w', w) = \begin{cases} 1 & \text{if } w' = w_y \\ 0 & \text{otherwise} \end{cases}$$

This process of deriving the new probability distribution P' from the original P is obtained by transfering the probability of every world w to its w_y, the most similar world to w where y is true.

Now we are able to show that $P(y \rightarrow x) = P'(x)$ or, using a terminology more appropriate to highlight the imaging process on y, that:

$$P(y \rightarrow x) = P_y(x)$$

where $P_y(x)$ is the new probability distribution derived from P by imaging on y. The probability of the conditional is the probability of the consequent after Imaging on the antecedent. In fact, as reported in [Lew81]:

$$\begin{aligned} P(y \rightarrow x) &= \sum_W P(w)\ w(y \rightarrow x) \\ &= \sum_W P(w)\ w_y(x) \\ &= \sum_W P(w)\ (\sum_{W'} I(w', w)\ w'(x)) \\ &= \sum_{W'} (\sum_W P(w)\ I(w', w)\ w'(x)) \\ &= \sum_{W'} P'(w')\ w'(x) \\ &= P'(x) \\ &= P_y(x) \end{aligned}$$

In the next Section we will see how we can apply Imaging to IR.

3.3 Retrieval by Logical Imaging

Taking into consideration a possible worlds semantics, the most obvious way of applying Imaging to IR would be by considering a document as a possible world, regarding it as a set of propositions with associated truth values. This is the view taken originally by Van Rijsbergen [vR86] and followed by others. In this view we should evaluate the probability of the conditional $d \rightarrow q$ by computing a new probability distribution P_d by imaging on d over all the possible worlds, i.e. over all the possible document representations. According to the definition of Imaging we have:

$$P(d \rightarrow q) = P_d(q) = \sum_D P(d)\ d_d(q)$$

where

$$d_d(q) = \begin{cases} 1 & \text{if } q \text{ is true at } d_d \\ 0 & \text{otherwise} \end{cases}$$

and d_d is the closest document to d where d is true.

In order to apply this technique to IR there are a few problems to be solved. The first is related to the computational requirements of Imaging. We need to assign a probability to each document (world) and to define and use a similarity measure between them. The former problem can be solved by looking at classical IR techniques, e.g. [vR79]. The latter problem is much more difficult to solve. It is related to the interpretation of the possible worlds semantics when the event d in the conditional statement is also interpreted as world. There is a difficulty with this interpretation since it is unlikely that a document d could not be true in d itself. To deal with this difficulty one would have to make explicit the difference between a document as a fictive object existing in its own right and a partial description of such an object (as in [LvR93]). Rather than doing this I have adopted a different approach.

I consider the set of terms T (index terms or simply terms used in the document collection) as the set of possible worlds. According to this I consider a process of Imaging on d over all the possible term t in T. More formally:

$$P(d \rightarrow q) = P_d(q) = \sum_T P(t)\ t_d(q)$$

where

$$t_d(q) = \begin{cases} 1 & \text{if } q \text{ is true at } t_d \\ 0 & \text{otherwise} \end{cases}$$

and t_d is the closest term to t for which d is true.

The possible worlds semantics in the context of IR can now be interpreted without difficulty by considering a term represented by a set (a vector) of documents. This is the inverse of the representation technique most often used in IR where a document is represented as a set of features, namely terms (or index terms). Intuitively this can be understood as "if you want to know the meaning of a term then look at all the documents in which that

term occurs". This idea is not new in IR (see for example [AK92, QF93]) and it has been widely used for the evaluation of term–term similarity (see Section 3.4). Using this representation technique for terms, we consider a document d true in a term (world) t if the term t occurs in d. Moreover, using a measure of similarity among terms it is easy to determine the closest term t_d to t which occurs in the document d. Using the same interpretation I consider a query q true in a term (world) t if the term t occurs in q. The process of Imaging on d causes a transfer of probabilities from terms not occurring in the document d (i.e. for which the document d is not true) to terms occurring in it.

Similarly we can also evaluate $P(q \rightarrow d)$ by imaging on q:

$$P(q \rightarrow d) = P_q(d) = \sum_T P(t)\ t_q(d)$$

where

$$t_q(d) = \begin{cases} 1 & \text{if } d \text{ is true at } t_q \\ 0 & \text{otherwise} \end{cases}$$

and t_q is the closest term to t where q is true.

Here we consider a process of Imaging on q over each possible term t in T so that the probability initially assigned to each term moves from terms not occurring in the query q to terms occurring in the query q.

Nie showed in [Nie88] that the two conditionals $d \rightarrow q$ and $q \rightarrow d$ have a very interesting interpretation in the context of IR. The conditional $d \rightarrow q$ expresses the *exhaustivity* of the document to a query, i.e. how much of a document content is specified by the query content. In fact $d \rightarrow q$ is intuitively equivalent to $d \subseteq q$. The conditional $q \rightarrow d$, instead, expresses the *specificity* of a document to a query, i.e. how much of a query content is specified in the document content. In fact, $q \rightarrow d$ is intuitively equivalent to $q \subseteq d$. Nie proposed to combine the two measures to produce a *correspondence* measure between query and document. This measure should estimate the relevance of a document to a query. I intend to investigate this proposal in the future.

The application of the above technique to IR requires an appropriate measure of similarity and an appropriate probability distribution over the term space T. I will tackle this problem in Section 3.4.

In the following two sections I explain in more detail the RbLI model using a simple example.

3.3.1 Evaluation of P(d → q)

We assume a set of terms T with a probability distribution P which assigns to each term $t \in T$ a probability $P(t)$ so that $\sum P(t) = 1$. We also use the following notation:

$$t(x) = \begin{cases} 1 & \text{if } t \text{ occurs in } x \\ 0 & \text{otherwise} \end{cases}$$

We assume we have a document collection D, with $d_i \in D$, where the documents are represented by terms in the set T. Finally, we assume we have a query q also represented by terms in T. Then, as explained in the previous section, it is possible to evaluate the $P(d_i \rightarrow q)$ as:

$$\begin{aligned} P(d_i \rightarrow q) &= P_{d_i}(q) \\ &= \textstyle\sum_T P(t)\ t_{d_i}(q) \\ &= \textstyle\sum_{\{\forall t: t(q)=1\}} P_{d_i}(t) \\ &\textstyle\sum_T P(t)\ t_{d_i}(q) \end{aligned}$$

where t_{d_i} is the ter$\sum_{\{\forall t: t(q)=1\}} P_{d_i}(t)$o t which also occurs in d_i, and $P_{d_i}(t)$ is the new probability distribution over the set of terms appearing in d_i obtained by imaging on d_i.

The evaluation of $P(d_i \rightarrow q) = P_{d_i}(q)$ must be repeated for each document in the collection D and it is based on the initial probability distribution over the set of terms T and on the availability of a similarity measure enabling the evaluation of t_{d_i}.

For a practical example of this evaluation let us suppose we have a query q described by the terms t_1, t_4, and t_6. We would like to evaluate the probability of relevance of a document d_i described by terms t_1, t_5, and t_6. Assuming a vector notation, Table 1 reports the evaluation of $P(d_i \rightarrow q)$ by imaging on d_i as an estimate of the probability of relevance of the document d_i to the query q.

The evaluation process is the following:

t	$P(t)$	$t(d_i)$	t_{d_i}	$P_{d_i}(t)$	$t(q)$	$P_{d_i}(t(q))$
1	0.2	1	1	0.3	1	0.3
2	0.1	0	1	0	0	0
3	0.05	0	5	0	0	0
4	0.2	0	5	0	1	0
5	0.3	1	5	0.55	0	0
6	0.15	1	6	0.15	1	0.15
$\sum_t$	1.0			1.0		0.45

Table 3.1: Evaluation of $P(d_i \rightarrow q)$ by imaging on d_i

1. Identify the terms occurring in the document d_i (third column of the table).

2. Determine for each term in T the t_{d_i}, i.e. the most similar term to t for which $t(d_i) = 1$. This is done using the similarity measure on the term space (fourth column).

3. Evaluate $P_{d_i}(t)$ by transferring the probabilities from terms not occurring in the document to terms occurring in it (fifth column).

4. Evaluate $t(q)$ for each term, i.e. determine if the term occurs in the query (sixth column).

5. Evaluate the probabilities $P_{d_i}(t(q))$ for all the terms in the query (seventh column) and evaluate $P_{d_i}(q)$ by summation (bottom of seventh column).

It is interesting to see a graphical interpretation of this process. In Figure 3.1(a) each term is represented by a world with its probability measure expressing the importance of the term in the term space T. The shadowed terms occur in document d_i. We assume a measure of similarity on the term space. Using this information we can now transfer the probability from each term not occurring in the document d_i to its most similar one occurring in d_i as depicted in Figure 3.1(b). In Figure 3.1(c) the terms with null probability disappear and those occurring in the query q are taken into consideration and their new probabilities $P'(t_i)$ are summed up to evaluate $P_{d_i}(q)$.

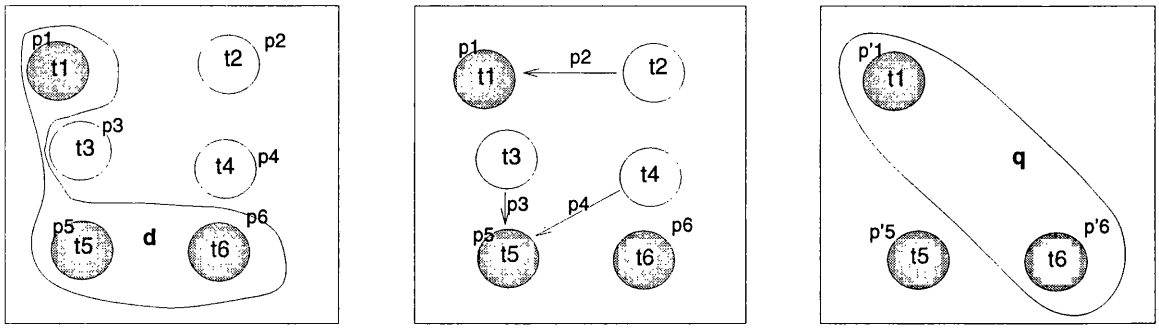

Figure 3.1: Graphical interpretation of the evaluation of $P(d_i \rightarrow q)$ by imaging on d_i.

3.3.2 Evaluation of P(q → d)

Using the same data of the previous example we can now evaluate the probability $P(q \rightarrow d_i)$. The terminology is analogous to that of the example above, though modified to take into consideration the evaluation of different elements. Table 2 reports the evaluation of $P(q \rightarrow d_i)$ which can be structured in the following steps:

1. Identify the terms occurring in the query q (third column of the table).
2. Determine for each term in T the t_q, i.e. the most similar term to t for which $t(q) = 1$ (fourth column).
3. Evaluate $P_q(t)$ by transferring the probabilities from terms not occurring in the query to terms occurring in it (fifth column).
4. Evaluate $t(d_i)$ for each term, i.e. determine if the term occurs in the document (sixth column).
5. Evaluate $P_q(t(d_i))$ for each term in the document and evaluate $P_q(d_i)$ by summation (seventh column).

A graphical interpretation of the Imaging process in relation to this example is reported in Figure 3.2.

t	$P(t)$	$t(q)$	t_q	$P_q(t)$	$t(d_i)$	$P_q(t(d_i))$
1	0.2	1	1	0.35	1	0.35
2	0.1	0	1	0	0	0
3	0.05	0	1	0	0	0
4	0.2	1	4	0.5	0	0
5	0.3	0	4	0	1	0
6	0.15	1	6	0.15	1	0.15
$\sum_t$	1.0			1.0		0.5

Table 3.2: Evaluation of $P(q \rightarrow d_i)$ by imaging on q.

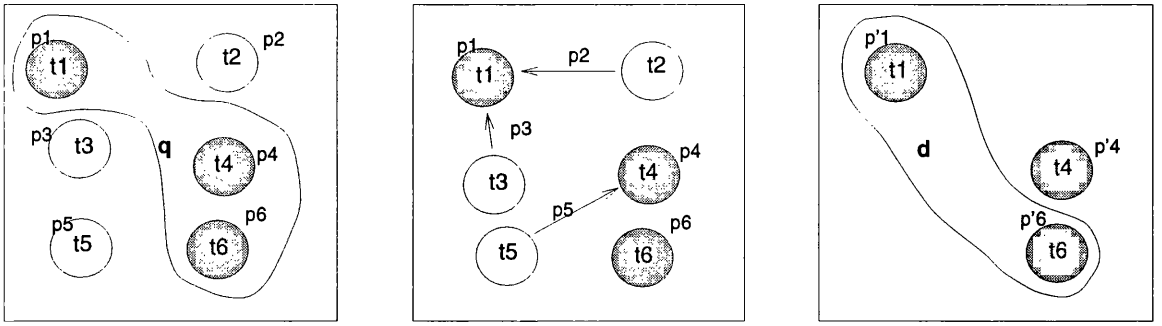

Figure 3.2: Graphical interpretation of the evaluation of $P(q \rightarrow d_i)$ by imaging on q.

3.4 Worlds mass and worlds distance

In order to perform RbLI we have two requirements:

- a probability distribution over the set of worlds which should reflect the importance of each world (the mass if we take an analogy with planets and stars) in the universe;
- a measure of similarity (which is related to distance) between worlds.

According to the view that a term is a world, these two requirements become: a probability distribution and a measure of similarity on the term space T.

The problem of determining an appropriate prior probability distribution over the set of index terms is one of the oldest problems of IR and many ways have been proposed for this purpose. The problem could be translated into finding a measure of the importance of a term in the term space, where this importance is related to the ability of the term to discriminate between relevant and not relevant documents. The importance of the term in the term space seems a reasonable rationale for a probability function. Several discrimination measures have been proposed, and a few examples can be found in [vR79, RS76]. For the tests reported in this chapter I used the *Inverse Document Frequency*, a measure which assigns high discrimination power to terms with low and medium collection frequency. Strictly speaking, this is not a probability measure since $\sum_t idf(t) \neq 1$, however we can assume it to be monotone to $P(t)$. We can use this estimate because we require only a ranking of the documents in response to a query, not the exact probability values.

The problem of defining a measure of similarity between terms and the use of such a measure for defining the accessibility among worlds is more difficult, although it has been addressed by many researchers in the past, in the fields of IR [WCY93, Voo93, Sri92] and Natural Language Processing [CH89, BDPd+92]. It is very important to chose a good measure since much of RbLI depends on it. I decided to use the *Expected Mutual Information Measure* ($EMIM$), because it is a well accepted measure in Lexicography [CH89].

In Information Theory $EMIM(i, j)$ is often interpreted as a measure of the statistical information contained in t_i about t_j (or vice versa, it being a

symmetric measure). The EMIM measure is defined as follows:

$$EMIM(i,j) = \sum_{t_i,t_j} P(t_i \in d, t_j \in d)\ log \frac{P(t_i \in d, t_j \in d)}{P(t_i \in d)P(t_j \in d)}$$

where t_i and t_j are terms.

We can estimate $EMIM$ between two terms using the technique proposed by Van Rijsbergen in [vR79]. This technique makes use of co-occurrence data that can be derived by a statistical analysis of the term occurrences in the collection. Using $EMIM$ we can then evaluate for every term a ranking of all the other terms according to their decreasing level of similarity with it. We store this information in a file which is used at run-time to determine for every term t its closest term occurring in d, that is t_d.

In the next section I will compare the performance of RbLI with simple weighted retrieval and I will also show that it is possible to decrease the computational effort of RbLI by cutting down the number of probability transfers using some heuristics.

3.5 Evaluating Retrieval by Logical Imaging

All the experiments reported in this section refer to the evaluation of RbLI for $P(d \rightarrow q)$.

For the experiments reported in this chapter I used the *Cranfield 2* document collection (in particular the C1400I). This test collection was produced in the Cranfield Project [CMK66] in the sixties and it is one of the most used for comparative evaluations. The collection is made up of 1400 documents, and 225 queries with relevance assessments. The number of terms used in the collection is 2686, manually derived from the documents. These experiments should be seen as illustrative of the technique and an indication of whether further research might be worthwhile.

Figure 3.3 reports a performance comparison between RbLI and a Benchmark. The Benchmark uses the same weighting scheme of RbLI, i.e. the IDF, but does not perform the transfer of probabilities which is typical of Imaging. As can be seen, the performance of RbLI are slightly better than the Benchmark, although a statistical analysis shows that the difference is not significant.

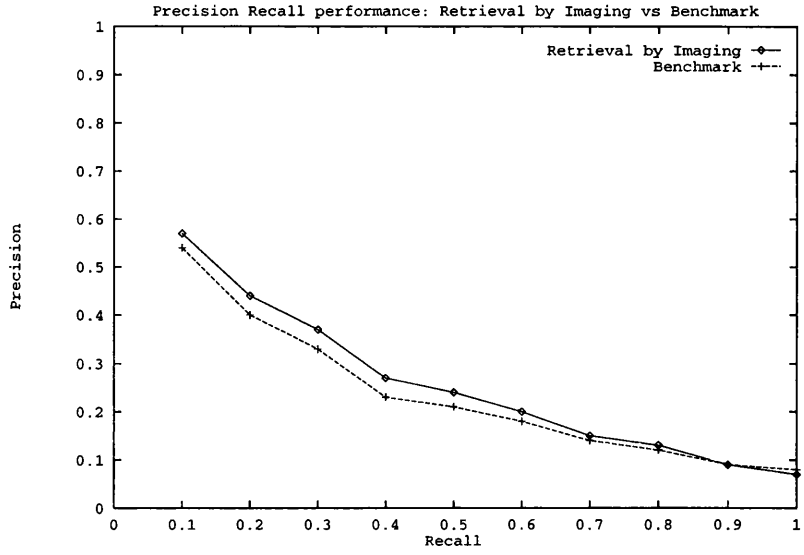

Figure 3.3: Performance of RbLI vs. Benchmark.

A problem with RbLI is the amount of computation necessary to provide for the transfer of probabilities. These computations need to be performed at run-time[2]. The next experiment investigates the possibility of reducing the amount of computation necessary at run-time to perform RbLI.

In Figure 3.4 I report the performance of RbLI when I reduce the number of probability shifts necessary to compute it. I decided to *cut off the* 10% *most frequent terms and the* 10% *least frequent terms* because their discrimination power is very low. During RbLI if a term t is not present in the similarity file because it was excluded from the similarity evaluation then we simply do not transfer its probability but we lose it. This is theoretically incorrect for the Imaging process since the new probability distribution P_{d_i} will not have $\sum P_{d_i} = 1$, but the ranking of the documents according to the estimates of $P(d \rightarrow q)$ does not change. The cut reduces considerably the amount of computation necessary at run-time and the results show that there is no significant decrease in performance. The size of the cut is, of course, dependent on the size and characteristics of the collection. The cut reported here is the biggest I was able to perform without decreasing the performance of RbLI for this collection.

[2]I am not concerned here with the computations necessary to the evaluation of EMIM between every pair of terms. This is certainly computationally very expensive, but it is performed off-line.

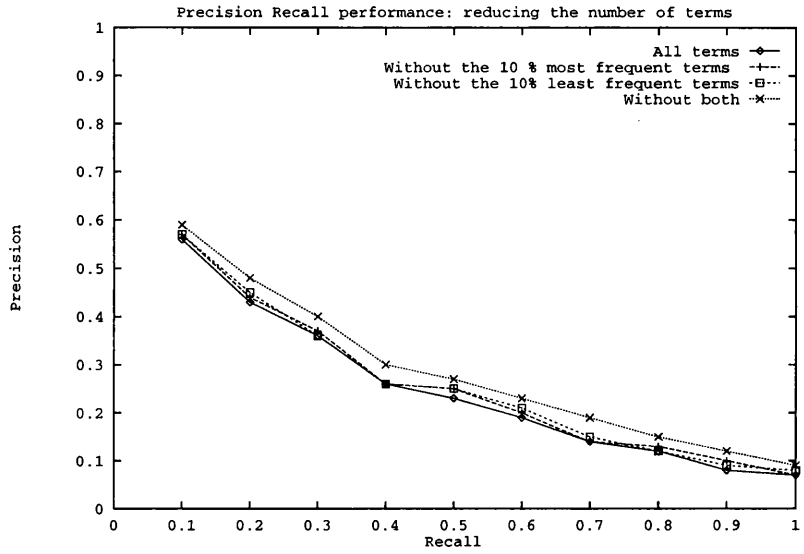

Figure 3.4: Performance of RbLI with different dimensions of the term space.

Another problem in the implementation of RbLI is related to the storage space requirements for the term similarity evaluated using EMIM (see Section 3.4). This is stored in a file which lists for every term all the other terms ordered by their similarity with it. This is used to evaluate for each term the closest one occurring in a particular document (or query). The dimension of this file can be considerably reduced if we store only *the first k most similar terms.* Again if for a term t we cannot find in the file its t_{d_i} then we do not transfer its probability but we lose it. This heuristics acts like a threshold on the accessibility between worlds. Figure 3.5 shows the performance of RbLI at various level of k. It can be seen that there is little difference between these values. For $k = 60$ there is actually a small increase in performance compared to RbLI without threshold. The use of a threshold on the similarity brings a considerable saving in storage space. For example, for $k = 60$ the file is reduced by almost 40%. It should be noticed that for $k = 0$ RbLI is equivalent to simple weighted retrieval because there is no probability transfer.

3.6 Related work

The use of imaging in IR was proposed for the first time by Van Rijsbergen in 1989 [vR89], however, to the best of my knowledge, there have been only a few attempts to use it.

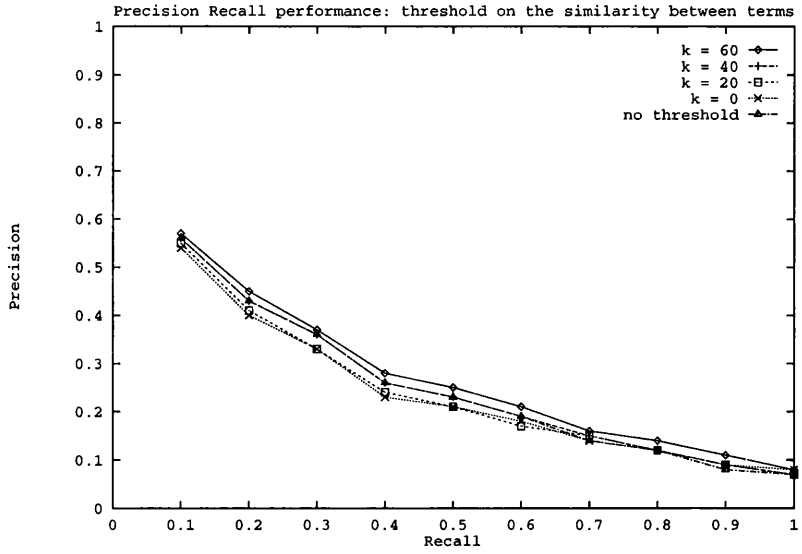

Figure 3.5: Performance of RbLI cutting the similarity measure between terms.

In [AK92] Amati and Kerpedjiev proposed two logical models for IR. One of them is based on conditional logic and makes use of imaging for the evaluation of $P(d \rightarrow q)$ and $P(q \rightarrow d)$. However, they proposed two different semantics for the evaluation of the two conditionals. For the evaluation of $P(d \rightarrow q)$ they consider a term as a world, while for the evaluation of $P(q \rightarrow d)$ they consider a document as a world. I see a difficulty in this latter approach because the event d in the conditional statement $q \rightarrow d$ is also interpreted as a world. To deal with this difficulty one would have to make explicit the difference between a document as a fictive object existing in its own right and a partial description of such an object (as in [LvR93]). Rather than do this I have adopted a different approach.

Sembok and Van Rijsbergen [Sv93] proposed a relevance feedback technique based on the use of imaging. Again, the perspective of a document as a world is used, which gives the same problem as before. Moreover, the similarity between documents is evaluated by means of clustering using nearest neighbours. The similarity measure used for the clustering on the document space is based on Dice's coefficient, a very simple similarity measure. I think that since most of the power of imaging relies on the correct identification of the closest possible world, it is very important to use the best possible similarity measure for the job.

Nie, first in [Nie92] and later in [NLB95], uses imaging to include in the retrieval process such contexts as user knowledge, domain knowledge, intentions, and so on. In his model both documents and queries are sentences. Possible worlds represent different states of the data set, for example possible states of knowledge that can be held by users. A document d is true in a world w if the document is "consistent" (the term is used here in a broad sense) with the state of knowledge associated with that world. Worlds differ because they represent different states of knowledge and, given a metric on the world space, we can identify the closest world to w for which d is true. Imaging can then be used for the evaluation of the certainty of the implication $d \rightarrow q$. Nie's approach takes a view similar to the one followed in this chapter. Both approaches consider a world as an informative entity, in the context of which a document or a query need to be checked for consistency. The major advantage of Nie's model is that it enables user modelling and therefore the evaluation of a user oriented measure of relevance, while RbLI only takes into account a system evaluated relevance.

There is also a number of papers that deal with techniques other than imaging for evaluating $P(d \rightarrow q)$. The work by Wong and Yao is perhaps the most interesting.

Wong and Yao [WY91, WY95] demonstrated that most of the IR models in use at present can be explained in terms of the formula $P(E \rightarrow H)$ that is evaluated as $P(H \mid E)$. The latter formula evaluates the degree of confirmation (or belief, according to the view taken) of the sentence H given evidence E. Conventional IR models can be obtained by associating either d or q to H or E, and by defining different ways of evaluating the probabilities via probabilistic inference on a concept space. Concepts are considered disjoint elements of the representation space, or are transformed in such a way to be disjoint. Terms are basic concepts.

Another important result is reported in [WY95] p. 58 where Wong and Yao show that their model, called "probabilistic inference model", subsumes the probabilistic model. Both Fuhr's probabilistic independence indexing model [Fuh89] and binary probabilistic independence retrieval model [vR79] can be explained in terms of the probabilistic inference model. Since the probabilistic inference model is based on the concept of conditional probability, then also Fuhr's and the binary probabilistic independence models are based on the same kinematics of probabilities. The amount of probability moved from one concept to another may change, but the principle remains the same: the transfer of probabilities provides the minimal revision of the prior probability that is necessary to make the evidence E certain without distorting the

profile of probability ratios on the representation space. Later in this thesis I will show that there are other types of probability kinematics that implement the Logical Uncertainty Principle in different ways. Moreover, the view taken by Wong and Yao is purely probabilistic. For them, only probabilistic inference is used for the evaluation of the uncertainty of the implication $E \rightarrow H$. I extend that view by taking into consideration a semantics of the representation space based on Possible World Semantics, which enables the evaluation of $P(E \rightarrow H)$ in a less restrictive way than does pure probabilistic inference. Thus I think that the use of the Possible World Semantics enables us to design and deal with different and more complex models of probability kinematics, like imaging.

I would also like to mention the work done in the context of expanding a query by adding terms that are semantically similar to those originally present. There are some similarities between my work on probability kinematics and work that has been done by others on query expansion. The transfer of probabilities that RbLI performs could be regarded as a way of expanding the terms present in the document with terms that are similar to them but not present. Work in this direction has been carried out by many researchers, for example Qiu and Frei [QF93] and Voorhees [Voo94]. However, the similarities between my work and query expansion are not easy to assess. There is nothing analogous to RbLI in the context of query expansion. Moreover, once we move out from the Possible World Semantics, it is very difficult to interpret the consequences of imaging in the context of natural language, while it is easy to interpret query expansion. The first results of a study of the implications of imaging in the context of sense resolution have appeared in [CSv96] and are reported in Chapter 5. Until the natural language semantics of imaging is fully understood, it is not possible to assess clearly the differences between imaging and query expansion or other query modification techniques.

3.7 Conclusions

In this chapter I have experimented with a new interpretation of the Imaging process for IR that I called RbLI. It is based on a possible worlds semantics where a term is a possible world. Every term (world) is assigned a probability and the accessibility between terms is measure by the EMIM. RbLI estimates the relevance of a document to a query using the probability of conditionals: either $P(d \rightarrow q)$ or $P(q \rightarrow d)$. I investigated RbLI for $P(d \rightarrow q)$ where this

is evaluated deriving a new probability on the term space by Imaging on the document d. The experiments reported here showed that RbLI is at least as effective as classical weighted retrieval on a small standard test collection and that the computational costs of its use can be considerably reduced using some simple heuristics.

Chapter 4

Probability Kinematics in Information Retrieval

In this chapter I analyse the kinematics of probabilistic term weights at retrieval time for different Information Retrieval models. I present four models based on different notions of probabilistic retrieval. Two of these models are based on classical probability theory and can be considered as archetypes of models long in use in Information Retrieval, like the Vector Space Model and the Probabilistic Model. The two other models are based on a logical technique of evaluating the probability of a conditional called imaging, one is a generalisation of the other. I analyse the transfer of probabilities occurring in the term space at retrieval time for these four models, compare their retrieval performance using classical test collections, and discuss the results. I believe that these results provide useful suggestions on how to improve existing probabilistic models of Information Retrieval by taking into consideration term-term similarity.

4.1 Introduction

In this chapter I explore further the use of the probability of a conditional, namely $P(d \rightarrow q)$, to estimate the conditional probability $P(R \mid q, d)$. I propose the use of a model called *Retrieval by General Logical Imaging*, based on a generalisation of the Retrieval by Logical Imaging model. I analyse and compare the probability kinematics of imaging and general imaging with that of two more classical probabilistic models: the Retrieval by Joint Probability

model and the Retrieval by Conditional Probability model. These two models are at the basis of many IR probabilistic models currently in use.

The chapter is structured as follows. Section 4.2 describes the representation model for documents and queries that I will use in the rest of the chapter. Section 4.3 describes the probability kinematics of four different retrieval models and explains in detail Retrieval by Logical Imaging and Retrieval by General Logical Imaging. The retrieval performance of these four models is then studied using the experimental settings described in Section 4.4 and the probability and similarity functions described in Section 4.5. The results are presented, compared, and discussed in Section 4.6. Section 4.7 reports the conclusions of the experimental investigation and the pitfalls in the proposed models.

4.2 The representation space

In probabilistic IR the task of the system can be formalised as follows. Let us assume binary relevance judgements, then $\mathcal{R}$, the set of possible relevance judgements, contains only the two possible events: relevance (R) and non-relevance ($\overline{R}$). The Probability Ranking Principle tells us that the task of a probabilistic IR system should be to rank documents according to their probability of being relevant: $P(R \mid \underline{q}, \underline{d})$, where $\underline{q}$ and $\underline{d}$ are the real query and the real document[1]. Unfortunately we can only estimate this probability by using the available query and documents representations, q and d. The probability $P(R \mid q, d)$, is then only an estimate of $P(R \mid \underline{q}, \underline{d})$ that depends very much on the quality of the document and query representation and on the quality of the estimation process. $P(R \mid q, d)$ is the probabilistic version of the *Retrieval Status Value* (RSV), a value assigned to each pair $(\underline{d}, \underline{q})$ that enables the ranking of all documents in the collection. The way the RSV is evaluated varies according to the IR model used.

The difficulty of applying probabilistic IR arises out of two different problems: *estimation* and *representation.*

The problem of *estimating* $P(R \mid q, d)$ is tackled in this chapter from a theoretical point of view. In the past researchers have tried to estimate $P(R \mid q, d)$ in many different ways. Often, researchers have had to resort to

[1]For "real query" and "real document" I mean the information need of the user and the informative content of the document. These are only ideal objects, but relevance can only be fully achieved with reference to these objects and not their representations.

ad hoc estimation techniques, very much dependent on experimental tuning of the parameters of their models. In Section 4.3 I will report on four theoretical retrieval models. Without entering into a discussion on the process of parameter estimation, I will analyse their differences and will draw some interesting conclusions that could be useful in directing new research in this area.

The effective *representation of documents and queries* is a very difficult problem. Most IR system assume a poor representation of documents and queries, based on the use of index terms automatically extracted from the text of documents and queries. In this chapter I will use the same poor representation technique, hoping that in the future more effective techniques will be available[2]. In fact, the novelty of the approach is in the assignment of a new semantics to this almost standard way of representing documents and queries.

4.2.1 Possible World Semantics and Logical Imaging

Possible World Semantics was introduced by Kripke [Kri71] in the context of Modal Logic. In this semantics the truth value of a logical sentence is evaluated in the context of a "*world*". The word "world" has been used like this by a number of logicians and seems to be the most convenient one, but perhaps some such phrase as "conceivable or envisageable state of affairs" (used in [HC68], p. 75) would convey the idea more clearly. Possible World Semantics has been used in Modal systems to give a semantics for Necessity (where a sentence is true in every possible world) and Possibility (where a sentence is true in at least one possible world)[3].

Without entering into the details of this semantics, one of the main advantages of Possible World Semantics is that it enables the evaluation of the truth value of a conditional sentence without explicitly defining the operator "$\rightarrow$" [Lew86]. What it requires is a clustering on the space of events (worlds) by means of a primitive relation of neighbourhood. The clustering then enables us to define an *accessibility relation* that is necessary for the evaluation of the conditional sentence. According to the Possible World Semantics the truth value of the conditional $y \rightarrow x$ in a world w is equivalent to the truth

[2]The retrieval techniques discussed in this chapter are independent of the representation technique used. We only assume that documents and queries are represented by means of relevant features.

[3]Here I simply refer to the Modal System $S5$ and not to more complex models.

value of the consequent x in the closest world w_y to w where the antecedent y is true [Sta81]. Ties at this stage, if they occur, are broken at random, to ensure the uniqueness of the closest world (but see further on for a generalisation). The passage from one world to another world can be regarded as a form of belief revision, and the passage from a world to its nearest neighbour is equivalent to the least drastic revision of one's beliefs. Using this process is a mean of implementing the logical uncertainty principle described in Chapter 1.

More formally, suppose we have a language L with an infinite set of propositional variables $\{a, b, c, ...\}$, two primitive connectives $\wedge$ (conjunction) and $\neg$ (negation), and parentheses. Suppose we have two sentences (well formed formulas) x and y of L. Moreover we have the additional connectives $\supset$ (material conditional), $\vee$ (disjunction), and $\equiv$ (material equivalence) defined in terms of the primitives.

We also assume we have a truth evaluation function τ that takes sentences into $\{0, 1\}$ and that meets the following two conditions:

(a) $\tau(\neg x) = 1 - \tau(x)$

(b) $\tau(x \wedge y) = \tau(x)\ \tau(y)$

Suppose now we have a finite set of possible worlds W. We can extend the truth evaluation function τ to indicate the truth value of a sentence in the context of a world:

$$\tau(w, y) = \begin{cases} 1 & \text{if } y \text{ is true at } w \\ 0 & \text{otherwise} \end{cases}$$

Let w_y be the world most similar to w where y is true. The implication $y \rightarrow x$ will be true at w if and only if x is true at w_y:

$$\tau(w, y \rightarrow x) =_{def} \tau(w_y, x)$$

This is the technique called *Logical Imaging* and was first proposed by Stalnaker [Sta81]. The arguments related to the existence of the most similar world to w are addressed in [Lew86, G$\ddot{\text{a}}$8]. I will not raise them here since I will be imposing a metric on W that will enable us always to find at least one most similar world to any given world w.

Imaging has been extended by Lewis [Lew81] to the case where there is a probability distribution on the worlds. Let us assume it follows the classical rules of probability, and in particular:

$$\sum_w P(w) = 1$$

Then we can go from probabilities of worlds to probabilities of sentences by summing the probabilities of the worlds where a sentence is true:

$$P(x) = \sum_w P(w)\ \tau(w, x)$$

This second probability distribution defined over the sentences is different from the probability distribution defined over the worlds, although the first can be derived from the second. However, for simplicity of notation we will use P for both.

Given a sentence y, we can derive a new probability distribution P' from the initial "prior" probability distribution P over the possible worlds:

$$P'(w') =_{def} \sum_w P(w)\ \sigma(w', w, y)$$

where:

$$\sigma(w', w, y) = \begin{cases} 1 & \text{if } w' = w_y \\ 0 & \text{otherwise} \end{cases}$$

The process of deriving the new probability distribution P' from the original P is obtained by transferring the probability of every "not-y world" w to its most similar "y-world". The new probability of the sentence x can again be evaluated as;

$$P'(x) = \sum_w P'(w)\ \tau(w, x)$$

Lewis showed that $P(y \rightarrow x) = P'(x)$ or, using a notation more appropriate to highlight the role of y:

$$P(y \rightarrow x) = P_y(x)$$

where $P_y(x)$ is the new probability distribution, called "posterior" probability, derived from P by imaging on y. In other words, the probability of the conditional is the probability of the consequent after imaging on the antecedent. The proof is reported in [Lew81]. The interested reader can also look at the following papers by Stalnaker [Sta81], Gärdenfors [Gä82] and Cross [Cro94] for more details of the imaging process.

In 1988 Gärdenfors proposed a generalisation of the imaging process [Gä88]. The generalisation originated from an attempt to overcome one of the restrictive assumptions Lewis made for Stalnaker's semantics of conditionals [Lew81]. The assumption is related to the "uniqueness" of the world w_y, that is the uniqueness of the most similar y-world to w. The generalisation that Gärdenfors proposed does not rely on this assumption[4]. The starting point is the use of a probability function $P^w(w')$ to represent the belief in the world w' given that the world w is certain. This probability function enables us to evaluate the (degenerated) probability function $P^w(y)$ that can be used to represent the fact that in any possible world w a sentence y can be true to a certain extent. The probability function $P^w(y)$ is derived from a probability distribution $P^w(w')$ over the possible worlds in such a way that:

$$P^w(y) = \sum_{w'} P^w(w')\tau(w', y)$$

Lewis called the probability function $P^w(y)$ "opinionated" because "it would represent the beliefs of someone who was absolutely certain that the world w was actual and who therefore held a firm opinion about every question" (see [Lew81], p. 145). Gärdenfors generalised imaging by considering the fact that, instead of having $P^w(y) = 1$ only for a single world w_y, we can have $0 \leq P^w(y) \leq 1$ for a set of worlds:

$$P^w(y) \begin{cases} > 0 & \text{if } y \text{ is true at } w \\ = 0 & \text{otherwise} \end{cases}$$

with the requirement that $\sum_w P^w(y) = 1$. Hence, taking into consideration the prior probability we go from probabilities of worlds to probabilities of sentences as follows:

$$P(y) = \sum_w P(w)\ P^w(y)$$

[4]He also characterised this generalisation of imaging in terms of a homomorphic condition that does not presuppose any kind of possible world semantics, but I will remain faithful to this semantics in the rest of this chapter.

From the prior probability distribution $P(w)$ we can derive a new probability distribution P'' so that:

$$P''(w') =_{def} \sum_{w} P(w)\ P^{w}(w')\ \sigma(w', w, y)$$

where:

$$\sigma(w', w, y) = \begin{cases} 1 & \text{if } w' \in W_y \\ 0 & \text{otherwise} \end{cases}$$

where W_y is the set of the closest worlds to w where y is true.

It could be proved, with a demonstration similar to the one reported in [Lew81], p. 142, that this new probability distribution is the posterior probability distribution derived from the prior probability P by *General Logical Imaging* on y. In other words, this new probability distribution can be obtained by transferring the probability from every world w to the worlds in W_y, the set of most similar (closest) worlds to w where y is true. The transfer of probability is performed according to the opinionated probability function $P^{w}(y)$. It is easy to prove that Lewis' imaging is just a special case of general imaging when $P^{w}(y) = 1$ for just one w.

The evaluation of $P_y(x)$ by general logical imaging is then performed in a similar way as the evaluation of $P_y(x)$ by logical imaging:

$$P_y(x) = \sum_{w} P''(w)\ \tau(w, x)$$

Again, it can be demonstrated that:

$$P(y \rightarrow x) = P_y(x)$$

The evaluation of $P_y(x)$ either by imaging or general imaging causes a shift of the original probability P from "not-y-worlds" to "y-worlds" to derive a new probability distribution P_y. Since the transfer of probabilities is directed towards the closest y-worlds, this technique is just what it is needed to implement the logical uncertainty principle described in Section 4.1. The probability revision is in fact minimal with regards to the accessibility relation, that is to say, it minimises the total distance covered in the probability transfer. In Sections 4.3.3 and 4.3.4 I will explain how we can use this result in the context of IR, but first let us examine how we can use the Possible World Semantics to model the term space.

	t_1	t_2	t_3	$\cdots$	t_n
d_1	1	1	0	$\cdots$	0
d_2	0	0	1	$\cdots$	1
d_3	1	1	1	$\cdots$	1
$\vdots$	$\vdots$	$\vdots$	$\cdots$	$\vdots$	$\vdots$
d_k	1	0	0	$\cdots$	1

Figure 4.1: The classical geometrical space semantics for the term space

4.2.2 The term space

One of the best known IR models is the Vector Space Model (VSM) [Sal68]. In the VSM a document is represented by means of a vector whose elements represent the presence/absence of certain features in the document representation, such as, for example, the presence or absence of index terms. Considering the binary case, for simplicity of exposition, a 1 in a particular position of the vector indicates the presence in the document representation of the feature associated with that position, while a 0 indicates its absence. The document representation space is therefore multidimensional, with as many dimensions as the number of features used for representing documents. Documents and queries are represented in this space as vectors. The semantics of the VSM is therefore that of a multidimensional geometrical space. The topology and the metrics of this space enable the evaluation of the RSV of a document with regards to a query as a distance measure. Many IR models use a similar representation space.

Here we use the same representation space but a different semantics. The semantics of the representation space is based on the Possible World Semantics. I use the Possible World Semantics in the context of IR by considering a term as a possible world, a view that was proposed in [CvR95]. According to this view, a term is represented as a "vector of documents". Intuitively this can be understood as "if you want to know the meaning of a term then look at all the documents in which that term occurs". This idea is not new in IR (see for example [QF93]) and it has been widely used for the evaluation of term–term similarity (see Section 4.5).

More formally, let us assume our representation space is made of a set of index terms T, we will call it a *Term Space.* The set of index terms T is our set of possible worlds. We also assume we have a document collection D where each document d is represented using terms in T, as depicted in the

	d_1	d_2	d_3	$\cdots$	d_k
t_1	1	0	1	$\cdots$	1
t_2	1	0	1	$\cdots$	0
t_3	0	1	1	$\cdots$	0
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\cdots$	$\vdots$
t_n	0	1	1	$\cdots$	1

Figure 4.2: Application of the Possible World Semantics to the term space

representation matrix in Figure 4.1. According to our semantics, in order to determine if a document is true or not in the context of a term it is sufficient to transpose the representation matrix and consider a document true in the context of a term if the document uses that term in its representation. The matrix depicted in Figure 4.2 can therefore be interpreted as representing the truth values of documents in the context of terms.

The above semantics for the term space can easily be extended to the case of a representation matrix with real values. In particular, if these values are in the $[0, 1]$ range, then they can be considered as probabilities of truth for a document in the context of a term.

In order to be able to apply imaging in this context, we also have to assume the presence of a prior probability distribution P on the term space, assigning to each term $t \in T$ a probability $P(t)$ so that $\sum_t P(t) = 1$. This probability reflects the importance of a term in the term space. We call this initial probability distribution "prior" because it reflects the importance of terms prior to the submission to the IR system of any query or the selection of any document as relevant to a user's information need. Once some external information enters the term space, mostly in the form of a query or a relevance judgement (but not necessarily only in these forms) then the importance of a term changes to reflect the new information. Accordingly, the probability assigned to a term changes to reflect the increased or decreased importance of the term. However for it to be considered a probabilistic space, the sum of the probabilities assigned to terms must remain constant (i.e. equal to 1) and so probabilities are moved around in the term space so that if one term increases its importance then some other terms must decrease their importance in a equal measure. These changes occurring in a IR system at retrieval time are very important in order to understand how IR models work. I believe that a study of the kinematics of probability in IR is very important in order to understand in detail why some models give better performance

than others. This is what I intend to investigate in the rest of the chapter.

4.3 Probability kinematics in IR

In the following sections I examine the different kinematics of probability that take place in four retrieval models. The purpose is to show how the probability associated with terms changes and shifts in different ways in different models as the result of new information entering the term space. I do not intend to associate directly any of the models presented here with existing IR models. However, these four models can be looked at as the archetypes of the most common IR models. In particular, the probability kinematics of the first two models, called Retrieval by Joint Probability and Retrieval by Conditional Probability, is similar to that taking place in the VSM and in the Probabilistic Retrieval model. In fact, apart from some normalisation factors, the VSM and the Probabilistic Retrieval model are respectively based on the concepts of joint probability and conditional probability. The last two models, called Retrieval by Logical Imaging and Retrieval by General Logical Imaging, are new and are based on a completely different approach for the transfer of probability. Their origin lies in the field of non-classical logics, and in particular in the application of the Logical Imaging technique to IR. I will show that in principle, i.e. without entering into complex "ad hoc" weighting and retrieval schemas, these last two models perform better than the first two. This result suggests that an improvement in retrieval effectiveness can be obtained by designing IR systems based on probabilistic models that use a non-classical probability kinematics.

In order to make the analysis clearer, I will provide examples of the kinematics of probability of the four models. I will take into consideration a particular document and query. We suppose we have a document d represented by terms t_1, t_5, and t_6 and a query q represented by t_1, t_4, and t_6. Each of these terms has a prior probability associated with it, indicated by $P(t)$. In the following I show how the RSV of document d is evaluated in different ways by different retrieval models and I concentrate my attention on how the probabilities associated with terms change and shift from term to term during the evaluation of the RSV. I indicate the new "posterior" probability with $P_d(t)$ to highlight the fact that it is obtained by looking at a particular document d.

4.3.1 Retrieval by Joint Probability

I call *Retrieval by Joint Probability* (RbJP) the ranking and retrieval of documents obtained by estimating the probability of relevance with the probability of the joint event of having both the query and the document true for a set of terms.

$$P(R \mid q, d) \approx P(q, d)$$

RbJP evaluates the RSV of a document using the following formula:

$$P(q, d) = \sum_t P(t)\ \tau(t, d)\ \tau(t, q)$$

Given a document d, we compute the sum of the probabilities of all terms that are present in both that document and that query. In Possible World Semantics this is equivalent to the sum of the probabilities of the worlds for which both the document and the query are true. It can easily be seen that here there is no transfer of probabilities. The prior probability $P(t)$ associated with term t does not change, but remains the same whatever document we are considering.

RbJP is the simplest approximation to $P(R \mid q, d)$, but it is used by many IR models. In fact it is the archetype of many IR models currently in use. Most IR models that are based on the evaluation of similarity between documents and query are based upon the idea of a joint probability measure. Both Dice's and Jaccard's coefficients (see [vR79], p. 39) are based on it, as can be seen once we remove the normalisation factors.

Let us consider, for example, the case $P(t) = k$ for every term in the term space, where k is a constant value. Then:

$$\begin{aligned} P(q, d) &= \textstyle\sum_t P(t)\ \tau(t, d)\ \tau(t, q) \\ &= k\ \ \| \, d \cap q \, \| \end{aligned}$$

where $\| \, S \, \|$ indicates the cardinality of the set S.

The results of the last formula is monotone to the *Retrieval by Simple Matching* value and with $k = 1$ we obtain the "coordination level coefficient" (the number of terms the query has in common with the document, see [vR79], p.

97) one of the oldest IR models. From the RbJP formula we can also obtain Dice's and Jaccard's coefficients just by assigning to k different normalisation factors. The *Cosine Correlation* used by the VSM is also a normalised version of RbJP, it is in fact the same formula with a Euclidean norm:

$$P(q,d) = \frac{\| d \cap q \|}{\| d \| \| q \|}$$

Moreover, let us suppose that $P(t)$ is estimated using the "Inverse Document Frequency" (idf) of the term t, defined as:

$$idf(t) = -log \frac{n}{N}$$

where n is the number of documents in which t occurs, and N is the number of documents in the collection. Let us also suppose that $\tau(t,d) = tf_d(t)$, and $\tau(t,q) = tf_q(t)$, where $tf_d(t)$ and $tf_q(t)$ indicate respectively the frequency of occurrence of term t in the document and in the query. Then we have the formula:

$$\begin{aligned} P(q,d) &= \textstyle\sum_t P(t)\ \tau(t,d)\ \tau(t,q) \\ &= \textstyle\sum_t idf(t)\ tf_d(t)\ tf_q(t) \end{aligned}$$

or, if the term frequency of occurrence of term t in the query is not considered, since very often a term occur only once in the query, we have the result:

$$P(q,d) = \sum_t idf(t)\ tf_d(t)$$

This corresponds to the *Cosine Correlation* using the "$tf \cdot idf$" weighting scheme as defined in [SY73]. We should notice that $P(q,d)$ is no more a measure of probability, since it does not give values between 0 and 1, however this problem could be simply solved introducing a normalisation factor.

While not intending to undervalue the importance of normalisation factors, I wish to point out that the probability kinematics of all the IR models I have mentioned does not change once a normalisation factor is introduced; it substantially remains the same as that of RbJP.

To show how I evaluate the RSV in the case of RbJP, I refer to Table 4.1, where I report the evaluation of $P(q,d)$ for a particular document and query. The evaluation process is the following:

t	$P(t)$	$\tau(t,d)$	$\tau(t,q)$	$P(t) \cdot \tau(t,d) \cdot \tau(t,q)$
1	0.2	1	1	0.2
2	0.1	0	0	0
3	0.05	0	0	0
4	0.2	0	1	0
5	0.3	1	0	0
6	0.15	1	1	0.15
$\sum_t$	1			0.35

Table 4.1: Example of the evaluation of $P(q,d)$

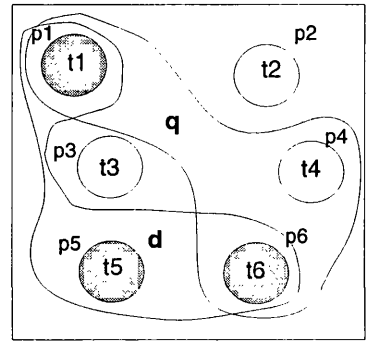

Figure 4.3: Graphical interpretation of the evaluation of $P(q,d)$.

1. Identify the terms occurring in the document d.
2. Identify the terms occurring in the query q.
3. Evaluate the $P(d,q)$ by summing the probability of all terms present in both document and query.

A graphical interpretation of RbJP using the Possible World Semantics is given in Figure 4.3, where each term is represented by a world with its prior probability measure expressing the importance of the term in the term space T. The shaded terms are those occurring in document d (Figure 4.3). The value of $P(q,d)$ is obtained by summing the probability of all terms occurring both in the document and in the query representations, that is summing the probabilities of the shaded terms also occurring in q.

This process in RbJP is not covered by the logical uncertainty principle, since there is no revision of the prior probability.

4.3.2 Retrieval by Conditional Probability

In the case of *Retrieval by Conditional Probability* (RbCP) the probability of relevance of a document is estimated by evaluating the conditional probability of the query given the document.

$$P(R \mid q, d) \approx P(q \mid d)$$

In other words, the relevance of a document is estimated by the extent to which the fact that we are observing that document supports the observation of the query. According to classical logics the conditionalisation, q given d is equivalent to the material implication $d \supset q$.

$P(q \mid d)$ can be evaluated as follows:

$$\begin{aligned} P(q \mid d) &= P_d(q) \\ &= \textstyle\sum_t P_d(t)\ \tau(t,q) \\ &= \textstyle\sum_t (1+\lambda_d)\ P(t)\ \tau(t,q) \end{aligned}$$

where $P_d(t)$ is the posterior probability distribution obtained by conditioning on the document d, and $(1+\lambda_d)$ is the factor by which the prior probability is to be modified to obtain the posterior probability. The value λ_d is the ratio between the sum of the probabilities of the terms not occurring in d and the sum of the probabilities of those that do occur in d:

$$\lambda_d = \frac{\sum_{t \not\in d} P(t)}{\sum_{t \in d} P(t)}$$

Notice that RbCP is a "normalised" form of RbJP. In fact, according to Probability Theory:

$$P(q \mid d) = \frac{P(q,d)}{P(d)}$$

The normalisation enables the prior probability to be revised in accordance with the characteristics of the particular document under consideration. Thus, the transfer of probabilities that takes place in RbCP provides the minimal revision of the prior probability necessary to make d certain without

t	$P(t)$	$\tau(t,d)$	$P_d(t)$	$\tau(t,q)$	$P_d(t) \cdot \tau(t,q)$
1	0.2	1	0.308	1	0.308
2	0.1	0	0	0	0
3	0.05	0	0	0	0
4	0.2	0	0	1	0
5	0.3	1	0.461	0	0
6	0.15	1	0.231	1	0.231
$\sum_t$	1		1		0.539

Table 4.2: Example of the evaluation of $P(q \mid d)$

distorting the profile of probability ratios. In fact, the posterior probability is proportional to the prior probability, so leaving constant the ratio of probabilities associated with the terms after the contraction of the term space due to the fact that the conditional event d has become certain.

Wong and Yao demonstrated in [WY95] that most probabilistic models of IR can be explained using $P(q \mid d)$. The difference between the various models is given by the different ways $P(q \mid d)$ can be evaluated.

Table 4.2 reports an evaluation of $P(q \mid d)$. The evaluation process is the following:

1. Identify the terms occurring in the document d.
2. Evaluate the posterior probability $P_d(t)$ by transferring the probabilities from terms not occurring in the document to terms occurring in it. The probabilities are transferred in a proportional way, so that each term occurring in the document d receives a portion of the sum of the probability of the terms not occurring in the document proportional to its prior probability.
3. Evaluate $\tau(t,q)$ for each term, i.e., determine the terms occurring in the query.
4. Evaluate $P_d(t) \cdot \tau(t,q)$ for all terms and evaluate $P_d(q)$ by summation.

It is interesting to see a graphical interpretation of probability kinematics induced by RbCP. In Figure 4.4(a) each term is represented by a world with its prior probability. The shaded terms occur in document d. The

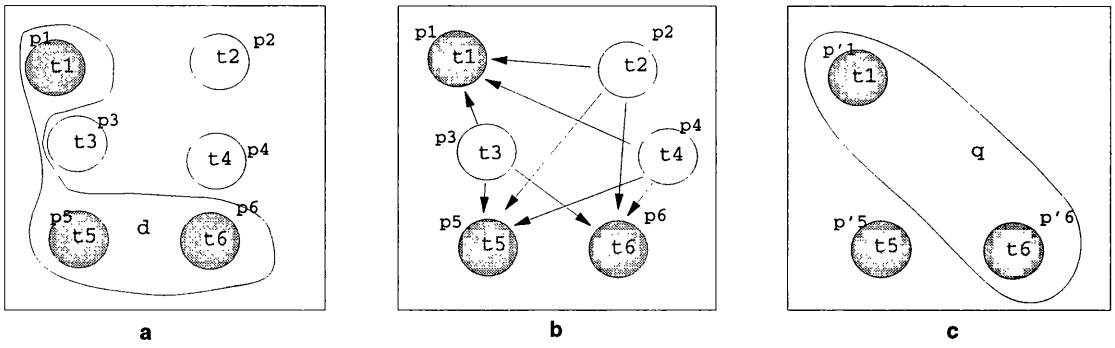

Figure 4.4: Graphical interpretation of the evaluation of $P(q \mid d)$.

conditioning process transfers the probability from terms not occurring in d to those occurring in it as depicted in Figure 4.4(b). In Figure 4.4(c) the terms with null probability disappear and only those terms occurring in the query q are taken into consideration to evaluate $P(q \mid d)$.

4.3.3 Retrieval by Logical Imaging

I use imaging in IR for estimating the probability of relevance of a document by means of the probability of the conditional $d \rightarrow q$:

$$P(R \mid q, d) \approx P(d \rightarrow q)$$

The motivation behind this approach is related to the underlying definition of Relevance (R). I accept a logical notion of relevance, in accordance with the work of Cooper [Coo71] and Van Rijsbergen [vR86]. Relevance is defined as the truth value of the implication $d \rightarrow q$. An equivalent interpretation of the truth of $d \rightarrow q$ is to consider d and q as events, then the "satisfaction" of a document described by d entails the satisfaction of a query described by q. The satisfaction of a document d means that the logical expression d is true in the current retrieval situation. A particular case in which d is true is when a document corresponding to d is retrieved, so a slightly narrower interpretation of the truth of $d \rightarrow q$ is: "the retrieval of d leads to the satisfaction of q". It is known in IR that relevance is often uncertain due to the uncertainty in the description of the contents of documents and queries, therefore we cannot talk about the truth of $d \rightarrow q$, but of the degree of certainty (or uncertainty) of the truth value. This leads us to talk about $P(d \rightarrow q)$.

Wong and Yao [WY95] suggested estimating $P(d \rightarrow q)$ by $P(q \mid d)$. The limitations of this approach are known in the area of logics by the name of "triviality results", and were well illustrated by Lewis in [Lew81]. According to these results $P(d \rightarrow q)$ and $P(q \mid d)$ would be equal only in certain extreme cases that are so simple that they can be considered "trivial"[5]. These results excluded that conditional probabilities could be used as a probabilistic logic dealing with conditionals. As a consequence, Lewis suggested estimating the probability of a conditional using Logical Imaging.

Retrieval by Logical Imaging (RbLI) is the model that estimates $P(R \mid q, d)$ by $P(d \rightarrow q)$, where the latter is evaluated using logical imaging. A detailed explanation of the RbLI model can be found in Chapter 3.

The application of Possible World Semantics on the term space enables us to apply imaging to derive the posterior probability $P_d(t)$ by imaging on d over all the possible terms (possible worlds) t in T. Probabilities are transferred according to the kinematics induced by the imaging process. More formally $P(d \rightarrow q)$ can be evaluated in the following way:

$$\begin{aligned} P(d \rightarrow q) &= P_d(q) \\ &= \textstyle\sum_t P_d(t)\ \tau(t, q) \\ &= \textstyle\sum_t P(t)\ \tau(t_d, q) \end{aligned}$$

where t_d is the closest term to t for which d is true, or in other words, the most similar term to t that also occurs in the document d. The application of imaging to IR requires an appropriate measure of similarity over the term space to enable the identification of t_d. This is the equivalent of the accessibility relation described by Lewis in [Lew81]. The measure of similarity used in the evaluations reported in this chapter is described in Section 4.5.

RbLI implements Van Rijsbergen's Logical Uncertainty Principle because it provides the minimal revision of the prior probability in the sense that it involves no gratuitous movement of probability from one world to dissimilar worlds. The revision of the prior probability necessary to make d certain is obtained by adopting the least drastic change in the probability space. This is achieved by transferring probabilities from each term not occurring in the document d to its closest (most similar) term occurring in d, so that the total amount of the distance covered in the transfer is minimal. A detailed

[5] Only a so called *trivial probability function*, according to which positive probabilities are never assigned to more than two incompatible alternatives, would accept the equivalence $P(d \rightarrow q) = P(q \mid d)$

t	$P(t)$	$\tau(t,d)$	t_d	$P_d(t)$	$\tau(t,q)$	$P_d(t) \cdot \tau(t,q)$
1	0.2	1	1	0.3	1	0.3
2	0.1	0	1	0	0	0
3	0.05	0	5	0	0	0
4	0.2	0	5	0	1	0
5	0.3	1	5	0.55	0	0
6	0.15	1	6	0.15	1	0.15
$\sum_t$	1.0			1.0		0.45

Table 4.3: Example of the evaluation of $P(d \rightarrow q)$ by imaging on d

comparison between other forms of conditionalisation and imaging can be found in [Cro94].

For a practical example of the evaluation of RbLI let us suppose we have the same query q and document d of the previous sections. Table 4.3 reports the evaluation of $P(d \rightarrow q)$ by imaging on d. The evaluation process is the following:

1. Identify the terms occurring in the document d.
2. Determine for each term in T the t_d, i.e. the most similar term to t for which $\tau(t,d) = 1$. This is done using a similarity measure on the term space not described here to keep the example simple.
3. Evaluate $P_d(t)$ by transferring the probabilities from terms not occurring in the document to terms occurring in it.
4. Evaluate $\tau(t,q)$ for each term, i.e. identify the terms occurring in the query.
5. Evaluate $P_d \cdot \tau(t,q)$ for all terms and evaluate $P_d(q)$ by summation.

A graphical interpretation of this process is depicted in Figure 4.5. I assume we have a measure of similarity on the term space. Using it we can transfer probability from each term not occurring in the document d to its most similar one occurring in d (Figure 4.5(b)). After the transfer of probabilities, terms with null probability disappear and those occurring in the query q are taken into consideration, so that their posterior probabilities $P_d(t)$ are added together to evaluate $P_d(q)$.

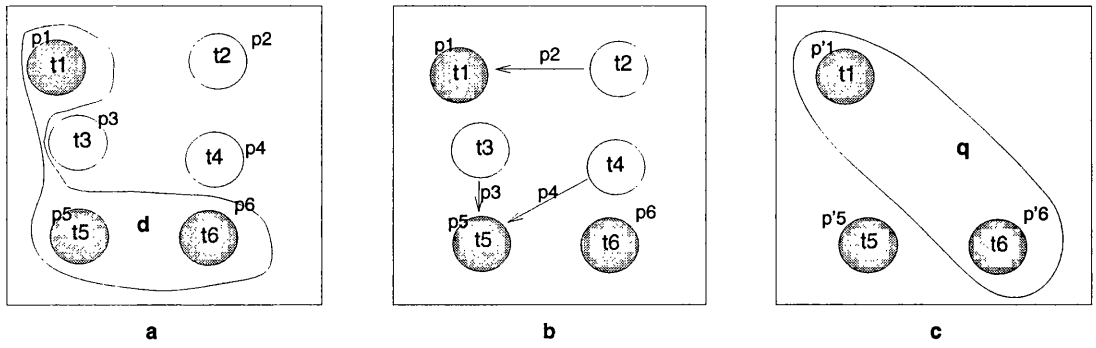

Figure 4.5: Graphical interpretation of the evaluation of $P(d \rightarrow q)$ by imaging on d.

4.3.4 Retrieval by General Logical Imaging

Retrieval by General Logical Imaging (RbGLI) is the result of the application of Lewis' general imaging technique to IR.

$$P(R \mid q, d) \approx P(d \rightarrow q)$$

In this case the evaluation of $P(d \rightarrow q)$ is performed using the following formula:

$$\begin{aligned} P(d \rightarrow q) &= P_d(q) \\ &= \textstyle\sum_t P_d(t)\ \tau(t, q) \\ &= \textstyle\sum_t P(t)\ (\sum_{t'} P^{t'}(d))\ \tau(t_d, q) \\ &= \textstyle\sum_t\ (\sum_{t'} P_d^{t'}(t)\ P(t'))\ \tau(t_d, q) \\ &= \textstyle\sum_{t,t'} P_d^{t'}(t)\ P(t')\ \tau(t_d, q) \end{aligned}$$

The opinionated probability function $P_d^{t'}(t)$, defined in Section 4.2.1, determines the amount of probability to be moved from t' to the term t belonging to the set T_d, where $T_d \subset T$ is the set of all terms occurring in document d. Such a function depends on the particular document on which general imaging is performed and on the particular term from which we want to transfer the probability. The number of opinionated probability functions required is then equal to the product of the number of documents multiplied by the number of terms. This number could be very high. I am currently working on this problem and I plan to use contextual information together with similarity information to determine the opinionated probability function, with the

t	$P(t)$	$\tau(t,d)$	t_d	$P_d(t)$	$\tau(t,q)$	$P_d(t) \cdot \tau(t,q)$
1	0.2	1	1	0.33	1	0.33
2	0.1	0	1; 6	0	0	0
3	0.05	0	5; 6	0	0	0
4	0.2	0	5; 1	0	1	0
5	0.3	1	5	0.47	0	0
6	0.15	1	6	0.2	1	0.2
$\sum_t$	1.0			1.0		0.53

Table 4.4: Example of the evaluation of $P(d \rightarrow q)$ by general imaging on d

document giving the context in which the similarity is measured. However, for the tests reported in this chapter I will make some strong assumptions:

1. The opinionated probability function is independent of the document being considered. This is equivalent to assuming that the opinionated probability function is context-independent, that is: $P_d^{t'}(t) = P^{t'}(t)$ for every $d \in D$.

2. The opinionated probability function does not use the similarity value, but only the similarity ranking. This means that in the evaluation of how much probability needs to be transferred from t' to t we will not consider the value of the similarity between t' and t, but only the position of t in a ranking of all terms in $(T - T_d)$ according to their similarity with t'.

These two assumptions enable us to use a single opinionated probability function for every term in the term space. I plan in the near future to remove first the second assumption and perform a transfer of probability that is related to the value of similarity between two terms, and later remove also the first assumption to make this transfer dependent on the context set by the document. In the tests reported in this chapter, however, I will make use of a very simplistic opinionated probability function. Such a function is described in detail in Section 4.5.

Table 4.4 reports an example of the evaluation of $P(d \rightarrow q)$ by general imaging on d. The evaluation process is the following:

1. Identify the terms occurring in the document d.

2. Determine for each term not occurring in the document (with $\tau(t,d) = 0$) the most similar terms (in this example only two terms) occurring in the document (those with $\tau(t,d) = 1$). This is done using a similarity measure on the term space.

3. Evaluate $P_d(t)$ by transferring the probabilities from terms not occurring in the document to terms occurring in it using the opinionated probability function . In this example the opinionated probability function prescribes that the most similar term to the one under consideration receives 2/3 of its probability, while the second most similar receives the remaining 1/3.

4. Evaluate $\tau(t,q)$ for each term, i.e. determine the terms occurring in the query.

5. Evaluate the probabilities $P_d \cdot \tau(t,q)$ for all the terms and evaluate $P_d(q)$ by summation.

A graphical interpretation of this process is shown in Figure 4.6. As can be seen in the picture, in the case of RbGLI the transfer of probability is performed from each term not occurring in the document d to the k_t most similar terms occurring in d. In the example $k_t = 2$ for every term, but k_t can be any other integer number so that $1 \leq k_t \leq l_t$, where l_t is the number of documents in which the term t occurs. The value of k_t is in theory different for every term, but can be set to a constant k independent of the term t. This setting simplifies considerably the evaluation of RbGLI. If $k_t = 1$ for every term, then RbGLI defaults down to RbLI. If $k_t = l_t$ for every term, then the transfer looks similar to the one produced by RbCP. However, note that the probability transfer in RbGLI is performed by taking into account the similarity between terms and not the ratio of prior probabilities as in the case of RbCP. Gärdenfors demonstrated in [Gä88] that it is not possible to find any prior probability distribution for which the transfer of probability induced by general imaging is equivalent to that induced by conditional probability.

4.4 Experimental analysis

So far I compared four probabilistic retrieval models using a common representation space and a common semantics. The comparison was mainly on theoretical grounds and was meant to show what happens at retrieval time to probabilities assigned to the elementary objects of the representation space.

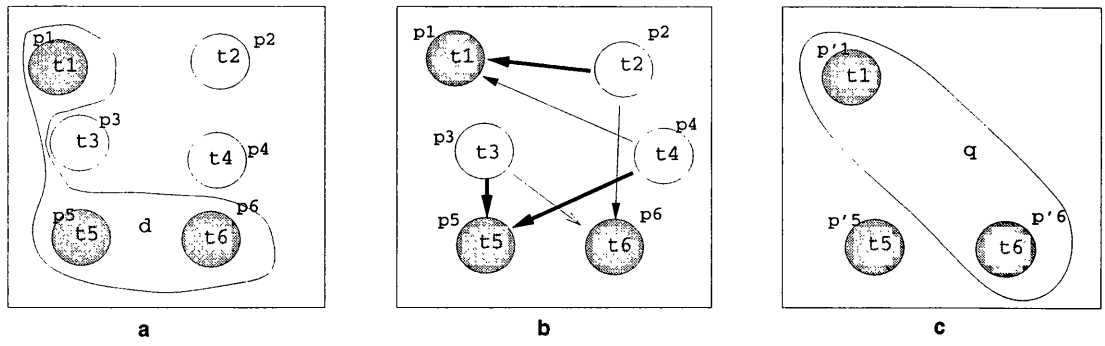

Figure 4.6: Graphical interpretation of the evaluation of $P(d \rightarrow q)$ by general imaging on d.

I showed that these four models induce different probability kinematics in the term space. Now, an obvious question comes to mind: which of the four models is the best in practical terms? The theoretical analysis I perform allows us to see the differences between the various models, but does not allow us to say which is the best. Some performance testing are necessary to compare the retrieval of the four models. These tests are only meant to show existing significant differences in the retrieval performance of the four different models. I decided to avoid using "ad hoc" indexing and normalisation schemes or adaptations of existing IR systems, since they could throw the comparison out of "balance". I used the retrieval models as they have been described in Section 4.3. My results cannot therefore be compared with the results achieved by other IR systems using the same data, and they only have a comparative significance in the framework of the present testing.

In order to study and compare the retrieval effectiveness of the four models under consideration I performed a series of tests using some standard test collections. I used three test collections that have been extensively studied and used in the field of IR: the *Cranfield 1400*, the *CACM*, and the *NPL* test collections. The characteristics of these three test collections are described in many papers (see for example [CMK66, Sv76]). A summary of the main characteristics is reported in Table 4.5.

The results of the test are presented using the standard evaluation technique used in IR. Precision and recall tables have been evaluated using their standard definition [vR79]. The method of linear interpolation has been used to determine the standard values corresponding to intervals of 10% in the recall figures.

Data	*Cranfield*	*CACM*	*NPL*
documents	1400	3204	11429
queries	225	52	93
terms in doc.	2686	7121	7492
terms in query	274	356	337
avg. doc. length	53.61	24.26	19.96
avg. query length	8.95	11.5	7.15
avg. rel. doc.	7.28	15.31	22.41

Table 4.5: Test collections data

4.5 Prior probability, similarity and opinionated probability function

In order to perform some performance testing with the four retrieval models, RbJP, RbCP, RbLI, and RbGLI, we have three requirements:

1. For all models, a "prior" probability distribution over the set of worlds which should reflect the importance of each world in the representation space;
2. For RbLI and RbGLI only, a measure of similarity (or a distance) between worlds;
3. For RbGLI only, an opinionated probability function.

The problems of determining the best prior probability and measure of similarity has already been discussed in Section 3.4. Here the same probability distribution (*idf*) and the same measure of similarity (EMIM) will be used.

The problems related to defining an appropriate *opinionated probability function* have already been introduced in Section 4.3.4. This third experimental requirement is a heavy one. The problem of finding a good opinionated probability function is still open. I do not tackle it in this chapter. In the tests reported in Section 4.6 I use a discrete monotonically decreasing transfer function that transfers from a term t a decreasing fraction of its probability to all the other terms in the term space once they are ordered in decreasing order of similarity. In particular, to simplify computations, in the evaluation of $P(d \rightarrow q)$ by general imaging on d, from each term not occurring in d

Average Precision values in % (increase in % over preceding model)				
Collection	RbJP	RbLI	RbCP	RbGLI
Cranfield 1400	24.3	27.6 (+12.0%)	31.8 (+13.3%)	36.2 (+12.1%)
CACM	27.1	33.2 (+16.8%)	37.1 (+10.6%)	42.8 (+13.4%)
NPL	22.4	29.8 (+24.8%)	38.1 (+21.9%)	42.1 (+9.5%)

Table 4.6: Comparison of the average precision of the four models with different test collections

the algorithm transfers probability only to the first 10 most similar terms occurring in d. Once terms are ordered in decreasing order of similarity with t, the probability transfer function I use works in such a way that the ith term gets double of what the $(i+1)$th gets. In the future I intend to use a more complex function that takes into account the contextual information provided by the particular document on which general imaging is performed. The opinionated probability function will be based on a term-term similarity measure evaluated in the context of that document.

4.6 Evaluation

I performed a comparative evaluation of the retrieval effectiveness of the four models presented above using the document collections and the experimental settings reported in Section 4.4 and 4.5.

I do not enter into the details of the evaluation, suffice to say that the actual computations of the RSV used for obtaining the figures reported in this section are very similar to the ones reported in the examples in Section 4.3. The only modifications to the techniques described earlier were introduced in order to reduce the number of computations necessary at run-time for the probability transfer. These modifications have already been described in detail in Chapter 3. Here I only compare the results obtained by each model to draw some plausible conclusions.

Table 4.6 reports the average precision values obtained by the different models on different test collections. It also shows the percentage increase in the average precision gained by using the different models. It can be easily seen that the average precision increases consistently from RbJP to RbGLI in all three document collections, although the increase rate is variable.

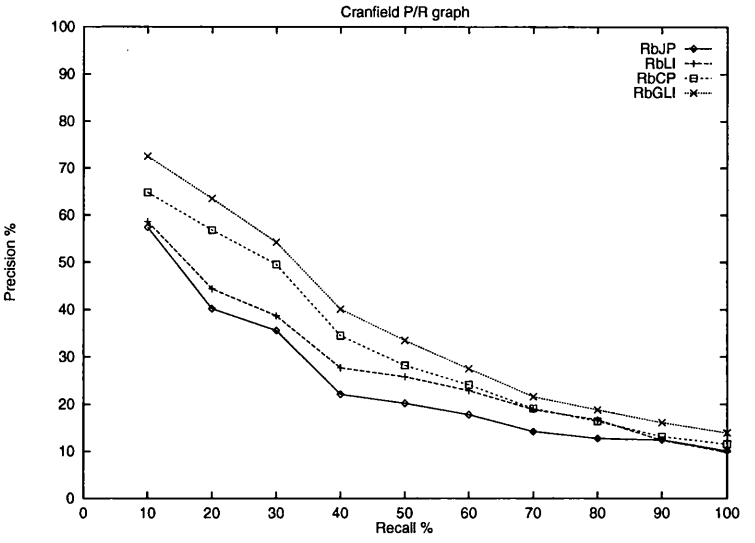

Figure 4.7: Precision and recall graphs for the Cranfield test collections

The results displayed in the Recall/Precision graphs in figures 4.7, 4.8 and 4.9 show that the performance of RbGLI are slightly higher than those obtained by any other model, with RbJP at the lowest level of performance.

From the results we can observe that:

- any model inducing a probability transfer (like RbLI, RbCP, and RbGLI) performs better than any model that does not induce such transfer (RbJP);
- any model that induces a probability transfer from one term to a set of terms (called "one-to-many" transfer, like in RbCP and RbGLI) performs better than any model in which either there is no transfer (RbJP) or the transfer is from one term to a single other term (called "one-to-one" transfer, like in RbLI);
- any model that induces a one-to-many transfer that takes into account the similarity between the donor and the receivers (RbGLI) performs better than any model with a one-to-many transfer that takes into account the probability ratio between the receivers (RbCP).

These findings are consistent over the three document collections.

Despite the fact that I am using a simple term weighting schema and that I am experimenting with small test collections, I think I can nonetheless con-

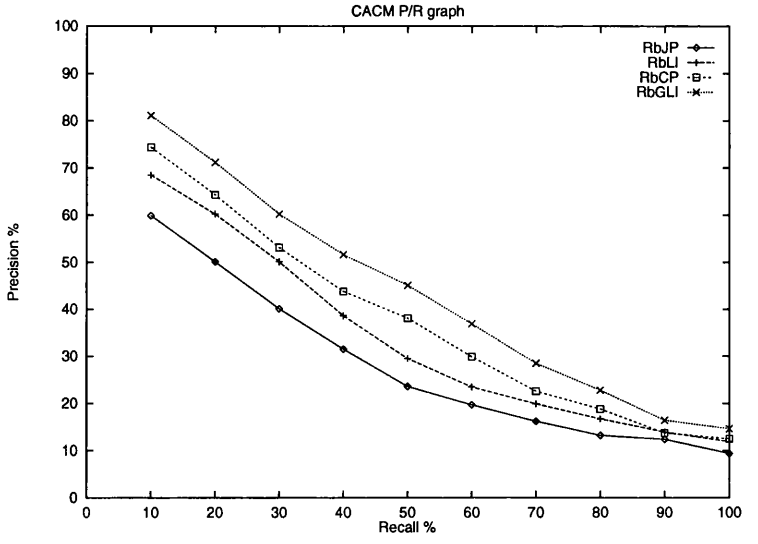

Figure 4.8: Precision and recall graph for the CACM test collection

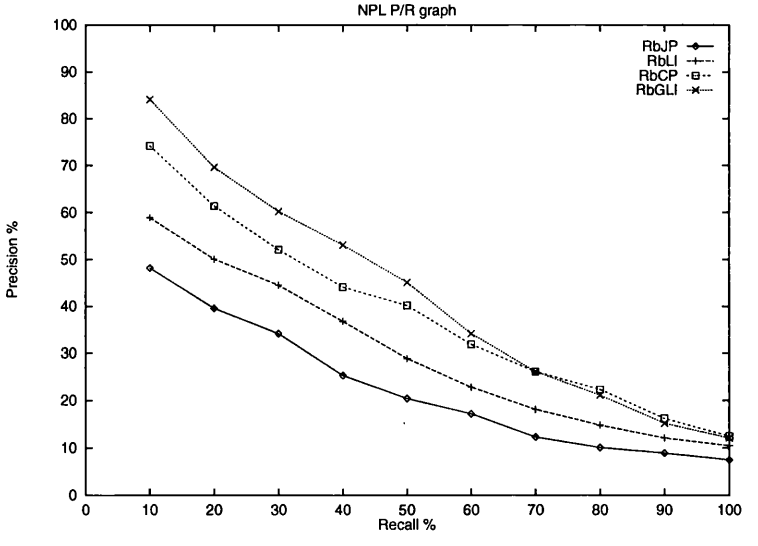

Figure 4.9: Precision and recall graph for the NPL test collection

clude that the probability kinematics of IR probabilistic retrieval models is worthy of further study. An exploration of the kinematics of probabilities in IR can help discover interesting properties of existing and new retrieval models, and can provide pointers for further study on how to improve the design of new IR models. An interesting result from this study on the kinematics of the four models presented is that it is possible to obtain higher levels of retrieval effectiveness by taking into consideration the similarity between the objects involved in the transfer of probability. However, the similarity information should not be use too drastically (like in RbLI) since similarity is often based on co-occurrence and such a source of similarity information is itself uncertain. A way of partially dealing with this latter uncertainty would be to contextualise the similarity information to make it document dependent. This is a line of research I will investigate using Natural Language Processing techniques. Some initial results of this work are reported in Chapter 5.

4.7 Conclusions

In this study of the probability kinematics in IR, I believe I have shown that, in principle, a probability transfer that takes into account a measure of similarity between the term "donor" and the term "recipient" is more effective in the context of IR than a probability transfer that does not take that into account. Most current probabilistic retrieval models are based on a probability kinematics that does not take into account similarity between terms or between documents, unless ad hoc weighting schemas, mostly based on clustering, are used. Furthermore, even when similarity between terms is taken into consideration, this is often just an add on to a conventional (rarely probabilistic) model, and it is not integrated into the model. I would therefore like to suggest a further investigation into more complex and optimised models for probabilistic retrieval, where probability kinematics follows non-classical models. General imaging is one of such models, but other ones can be developed using results achieved in other fields, such as non-classical Logics or Belief Revision theory.

The theory and the results reported in this chapter seem to suggest that an improvements in retrieval effectiveness can be obtained by designing IR systems that use probabilistic models based upon a different kind of probability kinematics. These results need to be tested experimentally using a larger collection of documents to see if they scale up.

Chapter 5

Sense resolution properties of Logical Imaging

In this chapter, the effect on word senses caused by Imaging is outlined and this is followed by a description of a small experiment and a proposal for further such experiments. Finally there is a short discussion and conclusions.

5.1 Word sense ambiguity

Before discussing the relationship between Imaging and word sense ambiguity, a brief overview of some of the features of ambiguity and disambiguation will be presented.

When ever dealing with words, it is important to remember that most words can refer to more than one sense. These individual senses can be quite distinct, for example the word "bat" can refer to an implement used in sports to hit balls or to a furry, flying mammal. Word senses can also be related, for example the word "crash" could refer to a physical event such as a car crash but also it could refer to the shares in a stock market dropping quickly. What sense a word has depends of course on the context that word appears in.

As information retrieval deals with the words of documents, and often ignores their context, inevitably IR is affected by word sense ambiguity. To illustrate, a manager of an on-line news retrieval system reported (in a personal communication) that the previous British Prime Minister was causing problems

with their retrieval system. A number of users had tried to retrieve articles about the Prime Minister using the query "major". This query caused many articles about "John Major" to be retrieved. However, in addition many more articles were retrieved where "major" was used as an adjective or as the name of a military rank.

5.1.1 Word sense disambiguation

The automatic disambiguation (or resolution) of word senses is a problem that has been studied for many years; Gale, Church and Yarowsky [GCY92] cite work dating back to 1950. These disambiguators were used in natural language processing applications such as translation systems. Early attempts to build disambiguators [Wei73, KS75, SR82] relied on a combination of hand built lexicons and rules. Although working well for the examples they were programmed for, researchers were never able to 'scale up' the disambiguators to work on large disambiguation problems.

However in the past ten years disambiguation research has moved towards investigating the exploitation of existing corpora already available online. The first work in this area was by Lesk [Les86] (an often cited paper). He used the textual definitions of an online dictionary to provide evidence for his disambiguator. The use of this evidence can be shown with a simplified example. Suppose we wish to resolve the sense of the word "ash" as it appears in the following sentence.

> *They cleared the* **ash** *from the coal fire.*

To disambiguate "ash", first its dictionary definition is looked up in the online dictionary and the individual senses of the word are identified. The format of the online dictionary is sufficiently structured to make this identification process relatively simple.

> *ash*: The soft grey powder that remains after something has been burnt.
>
> *ash*: A forest tree common in Britain.

Then the definitions of each of the other words in the sentence (apart from stop words) are looked up as well. For example:

coal: A black mineral which is dug from the earth, which can be burnt to give heat.

fire: The condition of burning; flames, light and great heat.

What follows is a process similar to ranked retrieval, where: the individual dictionary sense definitions of "ash" are regarded as a small collection of documents (a collection of two in this case); and the definitions of the words surrounding "ash" are regarded as the query. So the sense definitions are ranked by a scoring function that is based on the number of words co-occurring between a sense's definition and the definitions of each sentence word. The top ranked definition resulting from this process is chosen as the correct sense of "ash".

Lesk performed some limited testing of his technique and reported a disambiguation accuracy of between 50% and 70%. This level of accuracy is actually quite poor as Gale et al [GCY92] found that a disambiguator could have an accuracy of 75% if it always picked the most commonly occurring sense of a word. Although it is likely that if Lesk's disambiguator had incorporated information on the skewed frequency distribution of word senses its performance would have improved. However the importance of Lesk's work was to demonstrate the use of online corpora as sense disambiguation evidence and by doing this, to raise the possibility of building, without to much effort, a disambiguator capable of resolving the senses of a great many words.

Since Lesk's paper a bewildering array of disambiguators have been built using the same principle of collecting sense evidence from a large online corpus and ranking possible word senses according to the degree of match between sense evidence and the context of the ambiguous word: Cowie [CGG92], Wallis [Wal93] and Demetriou [Dem93] have made further use of dictionaries; Zernik [Zer91] built a disambiguator using a morphological analyser; Dagan [DIS91] used bilingual corpora; Church [Chu92] tried aligned bilingual corpora; Voorhees [Voo93] and Sussna [Sus93] used the WordNet thesaurus; and Yarowsky [Yar92] used a combination of Roget's thesaurus and Grollier's encyclopaedia to produce one of the better performing disambiguators to date, achieving 90% accuracy for the 12 words it was tested on.

A shared feature of all the disambiguators referred to above is an assumption that each individual sense of a certain word will appear in a wide context (typically 40–100 surrounding words) that is distinct from the contexts of the other senses of that word. It is not clear if this assumption is entirely correct

as research on human disambiguation has found that people can identify word senses accurately from a much narrower context of 1–5 words. This raises the possibility of having two senses of a word occurring in similar wide contexts but in different narrow contexts. Although such a situation probably accounts for some of the errors made by automatic disambiguators, when we consider that the Yarowsky disambiguator (cited above) makes this distinct context assumption and it has a 90% disambiguation accuracy, this assumption appears to be correct most of the time. It is this feature of distinct sense contexts coupled with the skewed frequency distribution of word senses (highlighted by Gale et al) that is important in the relationship between Imaging and the senses of a word.

5.2 Imaging and sense ambiguity

As has already been discussed, we can have two forms of Imaging in IR: Imaging on the document $P_d(q)$, and Imaging on the query $P_q(d)$. Each form behaves differently with regard to the senses of ambiguous words and will therefore be discussed separately. To illustrate these discussions, a simplified example will be used.

Let us imagine a document collection in which the word "bat" appears in a number of documents and that the frequency of occurrence of its word senses is skewed. In most documents, the word is used to refer to a sporting implement, but occasionally it is used to refer to the flying mammal. As the sporting sense of "bat" is predominant in this collection, words most similar to "bat" (similarity is measured by co-occurrence) will be those similar to this one sense. For this example, let us say that the words most similar to "bat" are "cricket", "baseball", "hit", and "ball". In terms of Imaging, it is these five words that are most likely to transfer their probabilities to each other.

Now let us look at two documents from this collection. Document d_1 is represented by words "bat" and "hit", while document d_2 is represented by words "bat" and "night". Document d_1 uses the word "bat" in the sporting sense (see Figure 5.1a); document d_2 uses it in the animal sense (see Figure 5.2a). Suppose a user enters the two word query, "bat", "cricket". How will the two forms of Imaging rank these two documents?

t	$P(t)$	$I(t, d_1)$	t_{d_1}	$P_{d_1}(t)$	$I(t, q)$	$P_{d_1}(t) \cdot I(t, q)$
bat	0.2	1	1	0.4	1	0.4
ball	0.1	0	5	0	0	0
night	0.05	0	1	0	0	0
cricket	0.2	0	5	0	1	0
hit	0.3	1	5	0.6	0	0
baseball	0.15	0	1	0	0	0
$\sum_t$	1.0			1.0		0.4

Table 5.1: Evaluation of $P(d_1 \rightarrow q)$ by Imaging on d_1

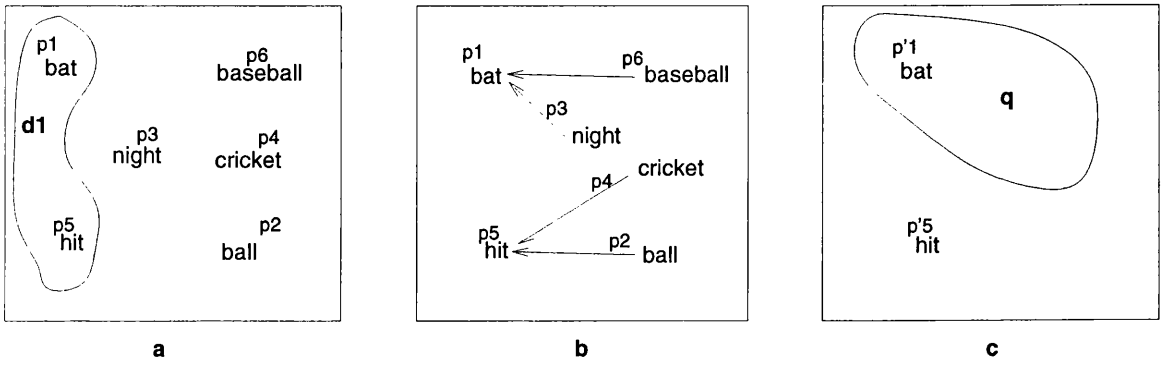

Figure 5.1: Sense resolution properties of $P(d_1 \rightarrow q)$ by Imaging on d_1.

5.2.1 Imaging on a document

As we recall, when Imaging on a document d, the probabilities of terms not appearing in d are transferred to the terms that do appear in d. The method of transfer is determined by a similarity measure which in this case is approximated using co-occurrence.

Looking at the example, let us first examine d_2. Since the words "cricket", "baseball", "hit", and "ball" are more similar to "bat" than to "night", all their probabilities transfer to this one word (Figure 5.2b). From Table 4 we can see that this transfer results in document d_2 having an estimated probability of relevance of 0.95.

However in the case of d_1, this document contains the word "hit". As this word is also similar to "cricket", "baseball", and "ball", the chances are that the probabilities of some of these words are likely to be transferred to "hit" instead of "bat", this is shown in Figure 5.1b. As "bat" is the only query

t	$P(t)$	$I(t, d_2)$	t_{d_2}	$P_{d_2}(t)$	$I(t, q)$	$P_{d_2}(t) \cdot I(t, q)$
bat	0.2	1	1	0.95	1	0.95
ball	0.1	0	1	0	0	0
night	0.05	1	3	0.05	0	0
cricket	0.2	0	1	0	1	0
hit	0.3	0	1	0	0	0
baseball	0.15	0	1	0	0	0
$\sum_t$	1.0			1.0		0.95

Table 5.2: Evaluation of $P(d_2 \rightarrow q)$ by Imaging on d_2.

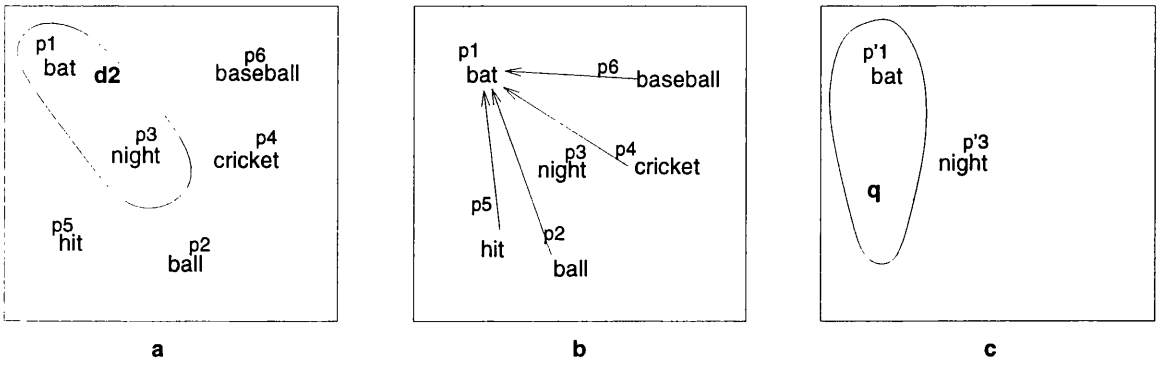

Figure 5.2: Sense resolution properties of $P(d_2 \rightarrow q)$ by Imaging on d_2.

word contained in d_1, this results in d_1 having a lower estimated probability of relevance than d_2 (see Table 3), which means that d_2 is ranked higher than d_1!

So what this example seems to show is that Imaging on a document will give preference to those documents which contain query terms appearing in unusual contexts. In terms of word senses, the supposition is that this form of Imaging will rank higher, those documents which hold query terms used in unusual senses.

5.2.2 Imaging on the query

When Imaging on a query, the method of probability transfer is similar to Imaging on documents except that the transfer is onto the terms in the query. Unlike Imaging on documents this form of Imaging is unaffected by the context in which query terms appear. From Figure 5.3 it can be seen

t	$P(t)$	$I(t,q)$	t_q	$P_q(t)$	$I(t,d_1)$	$P_q(t) \cdot I(t,d_1)$
bat	0.2	1	1	0.7	1	0.7
ball	0.1	0	4	0	0	0
night	0.05	0	1	0	0	0
cricket	0.2	1	4	0.3	0	0
hit	0.3	0	1	0	1	0
baseball	0.15	0	1	0.	0	0
$\sum_t$	1.0			1.0		0.7

Table 5.3: Evaluation of $P(q \rightarrow d_1)$ by Imaging on q.

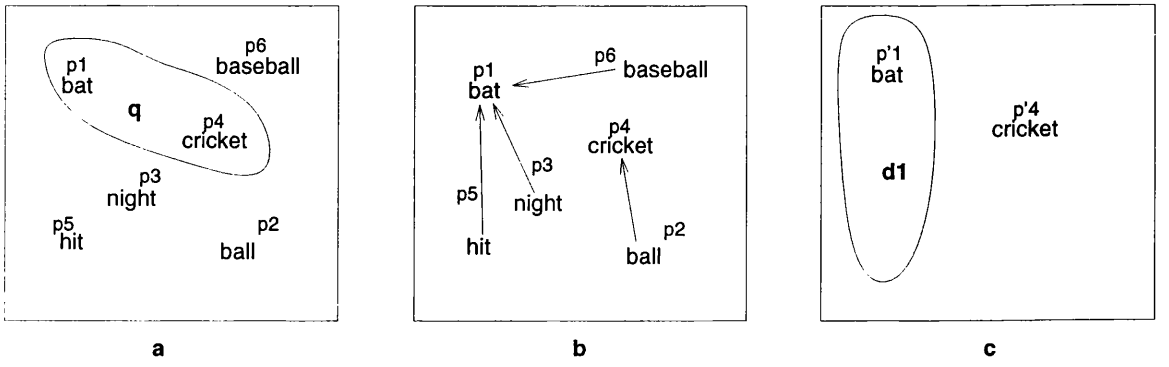

Figure 5.3: Sense resolution properties of $P(q \rightarrow d_1)$ by Imaging on q.

that the transfer of probabilities to the query terms is the same regardless of what document is being retrieved. Table 5 shows the estimated probability of relevance for d_1 and it is left as an exercise to the reader to show that d_2 will be assigned the same score.

5.3 Discussion and conclusions

The effect that Imaging on documents has on documents containing ambiguous query terms is caused because the Imaging technique is influenced by all the terms of a document and not just those that appear in the query. It is not clear whether this effect of preferring documents containing query terms in unusual senses or contexts is desirable. Term weighting schemes such as the popular $tf \cdot idf$ do give preference to unusual terms appearing in a document in unusually large quantities. Therefore one might think that this preference for the unusual might indicate that the Imaging effect is desirable. However

if a user enters a query term it would seem reasonable to expect him to intend the most common sense. Until the tests outlined in [CSv96] are completed though, I prefer to withhold judgement.

Chapter 6

Logical Imaging with Incomplete Knowledge of the Term Space

In this chapter I compare different ways of exploiting information about the Probabilistic Term Space used by a Probabilistic Information Retrieval System. These four models induce four different kinematics of probabilities, that use in different ways the information available about the Probabilistic Term Space. Some initial results show that the more information the Probabilistic Information Retrieval System uses in the retrieval process the better the system performs. A new model that exploit in a more complete way the information available to the system is then proposed. This model is particularly useful in the case of incomplete knowledge of the Term Space.

6.1 Introduction

A retrieval model based completely on the kinematics induced by General Logical Imaging is very computationally expensive, due to the large amount of probability transfers, has it has been shown in an experimental investigation using a very large document collection [CRSvR95]. In Section 6.4 I present a technique that put together the advantages of the models based on conditional probability and Imaging, and that partially reduces the computational burden of the Imaging (standard or general) process. In Section 6.5 I will show how to use this technique for dealing with incomplete knowledge

of the Term Space.

6.2 The probabilistic term space

Lets assume binary relevance judgements for documents; in other words the set $\mathcal{R}$ of possible relevance judgements contains only the two possible judgements: relevant (R) and not-relevant ($\overline{R}$). Then according to the Probability Ranking Principle [Rob77] the task of a probabilistic IR system is to rank the documents according to their probability of being relevant $P(R \mid \underline{q}, \underline{d})$, where $\underline{q}$ and $\underline{d}$ are the real query and the real document. Of course we can only estimate this probability by using the available query and document representations, q and d. The probability $P(R \mid q, d)$, is estimated using the *Retrieval Status Value* (RSV) that will be used by the IR system to rank documents. So:

$$P(R \mid \underline{q}, \underline{d}) \approx P(R \mid q, d) \approx RSV$$

Document and query representations are obtained by representing the query and the document informative content using descriptors available to the representation space of the IR system. Very often these descriptors are terms that are assigned manually or automatically to documents and queries. The representation space therefore corresponds very often with the *term space*.

One of the requirements of probabilistic IR is the existence of a "prior" probability distribution on the term space, assigning to each term a probability that is supposed to indicate the importance of the term in the term space. I will not discuss here how this prior probability is assigned to terms, I will just take it for granted. This probability distribution can be considered as a form of knowledge of the term space that is available to the IR system. It provides the IR system with information relative to the importance of a term in the term space. However, there is other knowledge of the term space that the IR system could acquire and use, and that is not explicit, but implicitly present in the term space. This last form of knowledge is relative to the semantic similarity between terms, that could be acquired, for example, through the use of a topology on the term space that can be induced by a thesaurus or by co-occurrence data. Some other form of knowledge about the importance of a term can be obtained. The user, for example could provide his own "subjective prior probability" or his own "subjective measure of similarity"

on the term space. These kinds of knowledge are very seldom used by the IR system in the evaluation of the RSV.

In the next Section I will show how the use of different forms of knowledge of the term space induces different probability kinematics in the evaluation of the RSV of a document with regards to a query.

6.3 Probability kinematics in IR

In the following sections I will examine the different kinematics of probability transfers that takes place in three different retrieval models. I will explain the changes in the probabilities of the term space by taking as an example a particular document d_i and a query q. In the representation space described in Section 6.2 let us suppose we have the document d_i described by terms t_1, t_5, and t_6 and the query q described by the terms t_1, t_4, and t_6. Each one of this term has a prior probability associated to it, these will be indicated by $P(t_i)$. In the following I will show how the RSV of the document d_i is evaluated by different retrieval models using the knowledge of the term space that is available to the IR system, and I will concentrate on how probabilities move from term to term during retrieval as a consequence of the use of a particular form of term space knowledge.

6.3.1 Retrieval without using term space knowledge

Some models of IR, like for example the models based on the evaluation of a similarity between document and query do not use term space knowledge. The only knowledge they use is related to presence or absence of a term in the document representation and in the query representation. In this case the RSV is evaluated simply by counting how many terms are present both in the query and in the document. There is no use either of the prior probability knowledge, or of the term similarity. The RSV is evaluated as some normalised form of the following formula[1]:

$$RSV = \| q \cap d \|$$

[1]This case in considered only theoretically in this chapter since it was shown long ago that IR systems based on a such a RSV perform quite badly.

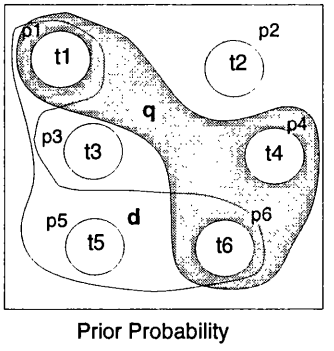

Figure 6.1: Graphical example of a model that does not perform probability transfer.

where $\| q \cap d \|$ is the number of term common to both the query q and the document d.

Some more advanced models of IR in this class of models do make use of some knowledge of the prior probability but only in oder to evaluate the RSV of a document as to the sum of the probabilities of the terms that are present in both the document and the query.

$$RSV = P(q, d)$$

where $P(q, d)$ is the probability of joint event q and d. This formula, like all the other formulas for evaluating the RSV, is often normalised to take into account the length of the document or of the query.

In these models there is no transfer of probabilities from term to term, and therefore no posterior probability, since no knowledge of the ratios of prior probabilities or of the similarity between terms is used (see Figure 6.1).

6.3.2 Retrieval using the term prior probability distribution knowledge

Most models of probabilistic IR are instead based on the concept of "conditional probability". Without entering into the details of these models and without considering normalisation factors often introduced in operative models, the probability kinematics that characterise this class of models is based on the preservation of the ratio of the prior probability. Probability are transfered in such a way that a term not present in a document moves its

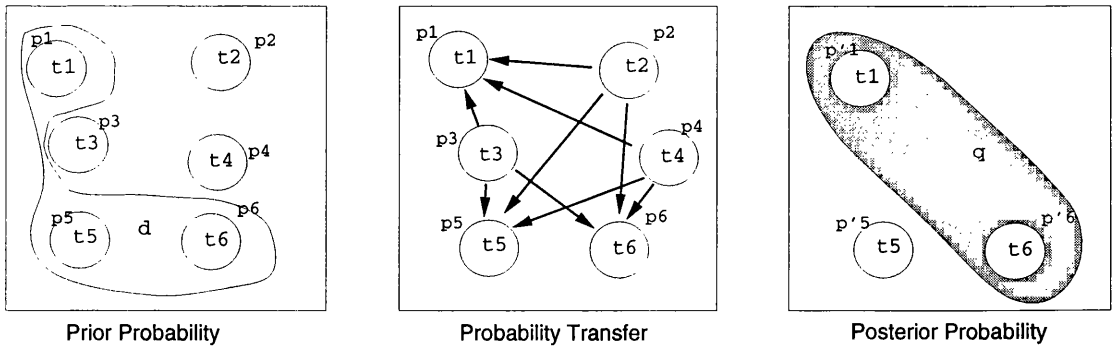

Figure 6.2: Graphical example of the use of prior probability distribution knowledge during probability transfer.

probability to all terms present in the document preserving the probability ratio among the terms present in the document (see Figure 6.2). In this way the posterior probabilities associated to these latter terms will be proportional to their prior probabilities. The RSV of a document is often evaluated with some normalised variation of the following formula:

$$RSV = P(q \mid d)$$

where $P(q \mid d)$ is the (Bayesian) conditional probability of the query q given the document d.

A disadvantage of this class of models is that they do not use some other important knowledge that is implicitly contained in the term space, like for example knowledge about the semantic closeness of terms. The only knowledge used is the one relative to the probabilistic importance of a term in the context of the term space.

6.3.3 Retrieval using term similarity knowledge

The two model of retrieval based on Imaging, the Retrieval by Logical Imaging model (RbLI) and the Retrieval by General Logical Imaging model (RbGLI), presented in [Cv95] use term-term semantic similarity to direct the transfer of probabilities from terms non present in the document to terms that are present. This enables the transfer of probability from a term not present in a document representation to its most similar other term that is instead present in the document representation, according to the Imaging process. I

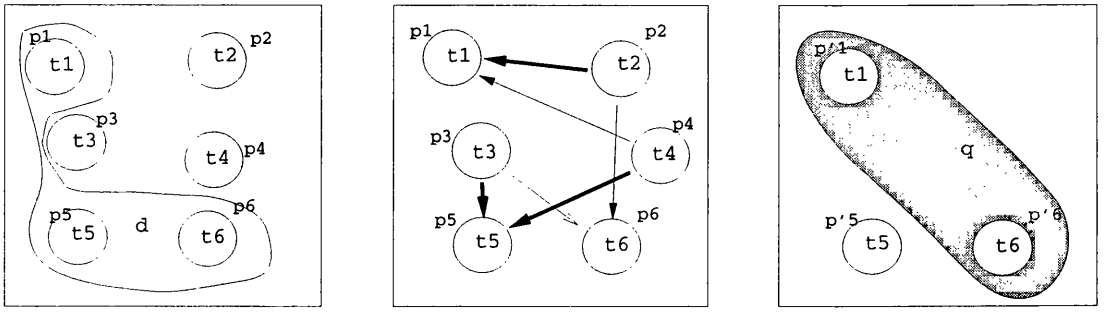

Figure 6.3: Graphical example of the use of term similarity knowledge during probability transfer.

will not enter into the complex details of the models of retrieval by (General) Imaging process here (see Chapters 3 and 4), it may suffice to indicate that the RSV is evaluated as follows:

$$RSV = P_d(q)$$

where $P_d(q)$ is the probability of the query q evaluated by (general) imaging on the document d.

The difference between the RbLI and RbGLI models is in the fact that RbLI transfers the probability of a term totally to its single closest term, while RbGLI transfers it to all terms present in the document with quantities that are in decreasing order in relation to the similarity between the "donor" and the "recipient" term. Figure 6.3 depicts an example of the probability kinematics induced by RbGLI. As it can be easily derived, RbGLI is a generalisation of RbLI.

The problem with both RbLI and RbGLI is that they require a very large amount of knowledge. It is in fact necessary to have a similarity value for every pair of terms in the term space. These values need to be used at retrieval time to find for every term not present in the document to the terms to which its probability needs to be transfered and the relative amount involved in the transfer. We should also remember that this computation needs to be done for every document in the document collection in order to produce a ranking according to their RSV.

6.4 Retrieval using both the term prior probability distribution knowledge and the term similarity knowledge

Experiments performed in Glasgow using a large collection of documents proved that models based on Imaging are very computationally expensive because of the complexity and number of the probability transfers involved. In trying to solve this problem I made some modification to the original Imaging models. The modifications partially followed some results already achieved using a small document collection and presented in Chapter 3. From these modification a new model was developed that not only enables to perform Retrieval by General Imaging in a faster and more efficient way but that also allows the combination of term prior probability distribution knowledge and term similarity knowledge.

This new model consists in performing the probability transfer according to term similarity knowledge only for a subset of all terms in the term space, that is only for those terms for which we are able to produce easily term similarity knowledge. The term similarity knowledge was produced using a measure called EMIM [vR79], an information theoretic measure based on term distribution information. I used only an estimate of this measure based on term co-occurrence data, thus on one hand enabling to produce term similarity knowledge in a efficient way, but on the other hand making it impossible to have similarity information on terms that are not co-occurring. The term similarity knowledge so produced and that is provided to the IR system is therefore not complete, but it is certainly the most useful part.

The probability transfer is then performed in the following way:

1. for the terms for which we have similarity knowledge the transfer of probability is done according to the RbGLI model;
2. for those terms for which we do not have similarity knowledge the probability transfer is performed according to the conditional probability paradigm, and therefore using the term prior distribution knowledge.

This combination of the Imaging and Conditional Probability enabled to use the most important part of the term similarity knowledge with regards to terms that have a high level of similarity with each other, while performing a transfer based on the fast and efficient term prior probability knowledge

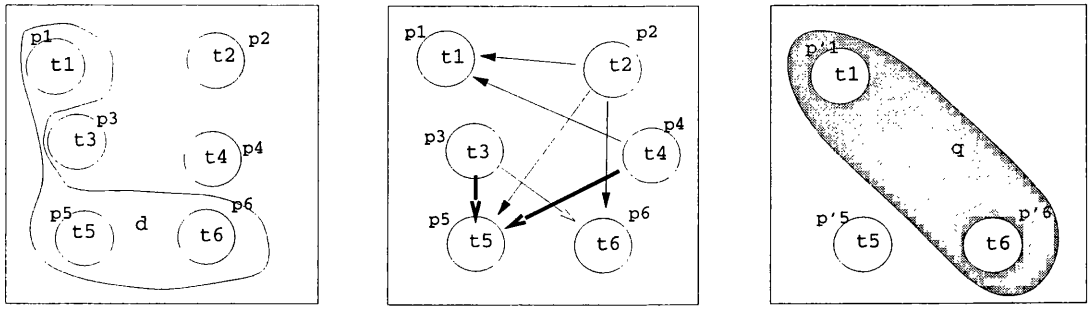

Figure 6.4: Graphical example of the use a combination of prior probability distribution knowledge and term similarity knowledge during probability transfer.

for all other terms. An example of this combination is given in Figure 6.4 where the probabilities of terms t_3 and t_4 are transfered according to RbGLI, while the probability of term t_2 is transfered according to the conditional probability paradigm.

This modification of Imaging enables us to combine and make use in a single model of both the term prior probability distribution knowledge and the term similarity knowledge.

6.5 Retrieval with incomplete knowledge of the Term Space

The technique presented in the previous section is very useful when it is impossible to have complete knowledge of the Term Space. This case is very frequent in practical applications of the RbGLI technique, in particular when large or dynamic collections are used. In the case of large collections it may be impossible to produce the amount of data necessary to have complete knowledge of the term similarity. This is the case we had to deal with in TREC-4 [CRSvR95] and that is reported in the Chapter 9.

Another very interesting case is related to "dynamic" collections, that is with applications that deals with collections that keep changing over time. This is the case of collections that are updated (adding or modifying documents) frequently. In this case, it is not possible to have complete knowledge of the term similarity, unless this is provided externally in the form of a quantitative

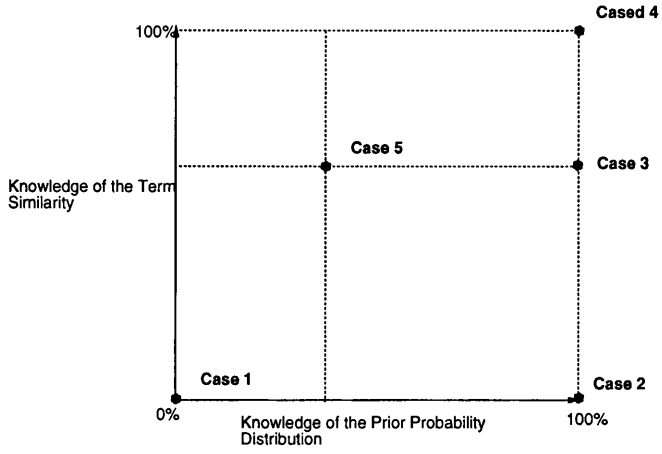

Figure 6.5: Levels of knowledge of the Term Space

thesaurus. Moreover, it is also impossible to have reliable knowledge of the term prior probability distribution. In order to have precise knowledge of the term prior probability distribution it would be necessary to re-index the collection every time a new document is added. Adding a new document not only changes the weights of terms already present in the Term Space, but can also add new terms to the space [BS74, vR79, VF95].

It is therefore not infrequent the case in which only incomplete knowledge of the term similarity and imprecise knowledge of the term prior probability distribution is available. Referring to Figure 6.5, it is possible to be in any of the following possible cases there depicted:

Case 1 no knowledge of the prior probability distribution and no knowledge of the term similarity. This is the simplest case, since there is no probabilities involved in the evaluation of the RSV of a document with regards to a query. In this case the RSV can be calculated using any of the IR models developed for the case of unweighted terms, like for example the Simple Matching Coefficient or the Dice's Coefficient models ([vR79], pp. 39).

Case 2 precise knowledge of the prior probability distribution and no knowledge of the term similarity. This is the case of most of the classical probabilistic IR models, where the knowledge of term similarity is not taken into account. Probability is transfered from term to term according to the RbCP model.

Case 3 precise knowledge of the prior probability distribution and incom-

plete knowledge of the term similarity. This is the case discussed in Section 6.4 and that is currently being tested using the TREC test collection [CRSvR95]. Probability transfer is performed as a combination of the RbGLI and the RbCP models.

Case 4 precise knowledge of the prior probability distribution and complete knowledge of the term similarity. This is the case of the application of the full RbLI or RbGLI models, and that has been tested in [Cv95]. Here we have complete and precise transfer of probabilities for the evaluation of the RSV.

Case 5 imprecise knowledge of the prior probability distribution and in complete knowledge of the term similarity. This is perhaps the most interesting and the most practically frequent case. How probabilities should be transfered in this case is currently being studied and tested.

6.6 Conclusions

In this chapter I outlined the probability kinematics of a new model that is based on the combination of the transfers induced by the use of prior probability distribution knowledge and term similarity knowledge. The model needs a deep theoretical study of the full consequences of the combination of these two kinds of knowledge to enable to take full advantage of both of them, in particular if this model needs to be used in the case of incomplete knowledge of the term similarity, imprecise knowledge of the term prior probability, or a combination of both.

Part IV

Implementation Study

Chapter 7

Logical Imaging and Probabilistic Datalog

Probabilistic Datalog, a probabilistic extension of the Datalog logical model of databases proposed by Fuhr in 1994 [Fuh95], enables the modelling of Information Retrieval as uncertain inference. The expressiveness of Probabilistic Datalog is such that it enables modelling both new models of hypermedia retrieval and classical probabilistic models of Information Retrieval.

In this chapter I report on some results and some open issues regarding the implementation of the General Imaging model on top of Probabilistic Datalog. This work was later carried on by Markus Blömer at Informativ VI, University of Dortmund, Germany, as part of his Diploma practical work. The results have been published in his Diploma Thesis [Blo97].

7.1 Information Retrieval by General Logical Imaging

Logical Imaging is a process developed in the framework of Modal Logic that enables the evaluation of a conditional sentence without explicitly defining the operator "$\rightarrow$" [Sta81]. Imaging has been extended to the case where there is a probability distribution on the worlds by Lewis [Lew81]. In this case the evaluation of $P(y \rightarrow x)$ causes a shift of the original probability P from a world w to the closest world w_y where y is true. Probability is neither created nor destroyed, it is simply moved from a "not-y-world" to a

"y-world" to derive a new probability distribution P_y. This process is called "deriving P_y from P by imaging on y".

General Logical Imaging originated from an attempt to overcome one of the restrictive assumptions Lewis made for Stalnaker's semantics of conditionals [Lew81]. The assumption is related to the "uniqueness" of the world w_y, that is the uniqueness of the world most similar to w where y is true. In [GÖ88] p. 110, Gärdenfors propose a generalisation of the Imaging process that does not rely on this assumption. The starting point of the generalisation is the use of a probability function to represent the fact that in any possible world w a proposition y is either true or false. In the case $P^w(y) = 1$ if y is true in w, and $P^w(y) = 0$ if y is false in w we go back to the classical definition of Imaging, in any other case, with $0 < P^w(y) < 1$, we state that y is only partially true in w. Lewis called such probability function "opinionated" because "it would represent the beliefs of someone who was absolutely certain that the world w was actual and who therefore held a firm opinion about every question" (see [Lew81], p. 145).

Retrieval by General Logical Imaging (RbGLI) can be regarded a the process of applying General Imaging on d in order to evaluate the probability that a document d implies the query q. The focus point of the RbGLI model, proposed by Crestani and Van Rijsbergen in [Cv95], is the consideration that "an index term is a world". The following is a simplified formulation of the model:

$$P(d \rightarrow q) \quad = \quad \sum_{t} \sum_{t'} P_d^{t'}(t) * \mu(t') * \tau(t, q).$$

where: $P_d^{t'}(t)$ is the opinionated probability of term t' in t in relation to document d; $\tau(t, q)$ is a function with the values 1 if the term t is present in the query q, and 0 if the term t is not present; and $\mu(t')$ is the probability assigned to each term in the term space T[1].

The application of the above technique to IR requires an appropriate measure of similarity over the term space T to enable the identification of the set of closest terms to t where d is true, and an opinionated probability function

[1]It should be noticed that the nature of the probability function $P(x)$ of $P(d \rightarrow q)$ is different from that of the probability function $\mu(x)$ used to assign the probability to index terms.

t_i	$P_d^{t_1}(t_i)$	$P_d^{t_2}(t_i)$	$P_d^{t_3}(t_i)$	$P_d^{t_4}(t_i)$	$P_d^{t_5}(t_i)$	$P_d^{t_6}(t_i)$
t_1	1.0	0.67	0	0.33	0	0
t_2	0	0	0	0	0	0
t_3	0	0	0	0	0	0
t_4	0	0	0	0	0	0
t_5	0	0	0.67	0.67	1.0	0
t_6	0	0.33	0.33	0	0	1.0

Table 7.1: An example of an opinionated probability function

t_i	$\mu(t_i)$	$\tau(t_i, d_1)$	t_d	$\mu_{d_1}(t_i)$	$\tau(t_i, q)$	$\mu_{d_1}(t_i) * \tau(t_i, q)$
t_1	0.2	1	t_1	0.333	1	0.333
t_2	0.1	0	t_1, t_6	0	0	0
t_3	0.05	0	t_5, t_6	0	0	0
t_4	0.2	0	t_5, t_1	0	1	0
t_5	0.3	1	t_5	0.467	0	0
t_6	0.15	1	t_6	0.2	1	0.2
$\sum_{t_i}$	1.0			1.0		0.533

Table 7.2: An example of the evaluation of $P(d_1 \rightarrow q)$ by general imaging on d

determining the portion of the probability associated to a term t that need to be transfer to a term t'. An example of a very simple opinionated probability function is reported in Table 7.1.

Table 7.2 reports an example of the evaluation of $P(d_1 \rightarrow q)$ by general imaging on d. A graphical interpretation of this process is depicted in Figure 7.1.

7.2 Probabilistic Datalog

Probabilistic Datalog [Fuh95] is an extension of stratified Datalog [Hul88]. The basic ideas of probabilistic Datalog are the assignment of probabilistic weights to facts and the computation of the weights of derived facts by means of intensional semantics.

Consider the following two uncertain facts:

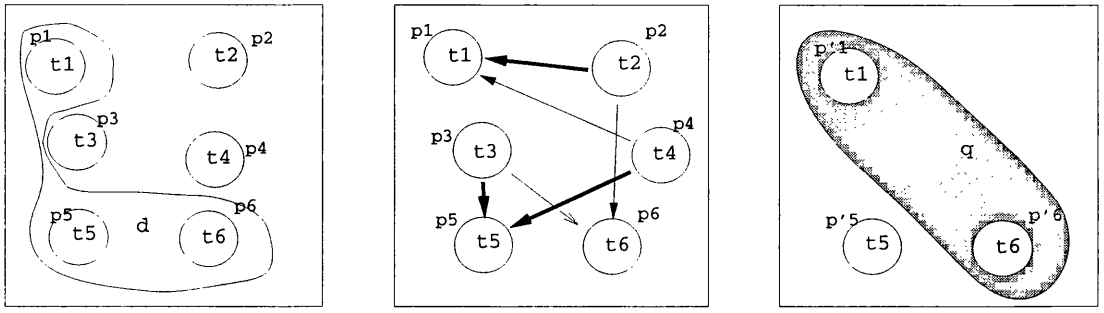

Figure 7.1: Graphical interpretation of the evaluation of $P(d_1 \rightarrow q)$ by general imaging on d.

0.7 docTerm(1,ir).
0.8 docTerm(1,db).

The first uncertain fact assigns the probability 0.7 to the proposition "document 1 provides information about the term ir". It is straightforward to regard *docTerm* as a probabilistic relation $\{(0.7(1, ir)), (0.8(1, db))\}$. To refer to the probability of a fact we write $\omega(docTerm(1,ir)) = 0.7$.

Given this little program, we may formulate a rule for deriving documents about both ir and db:

q1(D) :- docTerm(D,ir) & docTerm(D,db).

Assuming the tuples of *docTerm* to be stochastically independent, we get the relation $q1 = \{(0.56(1))\}$ where $\omega(q1(1)) = 0.56$. Rules enables the formulation of any conjunction and disjunction of predicates.

Now we extend our program with further facts and rules expressing the link structure of a hypertext document.

0.5 link(2,1).
about(D,T) :- docTerm(D,T).
about(D,T) :- link(D,D1) & about(D1,T).

The predicate *about* indicates the supposition that a document D is about term T, if it refers to a document that is about T.

We rewrite *q1* and get

q2(D) :- about(D,ir) & about(D,db).

An evaluation based on extensional semantics would yield $\omega(\text{q2(2)}) = 0.5 * 0.7 * 0.5 * 0.8 = 0.14$.

From a probabilistic point of view this weight is not correct, since *link(2,1)* is considered twice.

Instead of just computing the weight of a derived fact as a function of the rule weight and the weight produced by evaluating the rule body, probabilistic Datalog uses so-called event expressions for computing the resulting weight of a fact. Thus we achieve an intensional semantics of the fact weights. The central point is the extension of the relations with a special attribute for event expressions [FR95]. The event expression of a fact p is denoted by $\eta(p)$. The weight of a fact is computed via the formula $\omega(p) = P(\eta(p))$ where P is a probability distribution on events. Thus we get the probabilistically correct weight of *q2(2)*:

$$\begin{aligned}
\omega(q2(2)) &= P(\eta(q2(2))) \\
&= P(\eta(about(2,ir)\ \&\ about(2,db))) \\
&= P(\eta(about(2,ir)) \wedge \eta(about(2,db))) \\
&= P(\eta(link(2,1)\ \&\ docTerm(1,ir)) \wedge \\
&\qquad \eta(link(2,1)\ \&\ docTerm(2,db))) \\
&= P(link(2,1) \wedge docTerm(1,ir) \wedge \\
&\qquad link(2,1) \wedge docTerm(2,db)) \\
&= 0.5 * 0.7 * 0.8 = 0.28.
\end{aligned}$$

The example illustrates the computation of the event expression. Since P is a function on event expressions, it detects the double occurrence of the event *link(2,1)* and computes the correct probability.

To explain what we mean by the probability of a fact, we use possible worlds semantics [Nil86].

The probability of a fact $\varphi(x)$ is computed by summing the probabilities of those possible worlds, where the fact is true.

$$\omega(\varphi(x)) = \sum_{w_i \in \mathcal{W}} \mu(w_i) * \begin{cases} 1, & \text{if } \varphi(x) \text{ is true at world } w_i \\ 0, & \text{if } \varphi(x) \text{ is not true at world } w_i \end{cases}$$

Probabilistic Datalog can also cope with a special case of dependent events, namely disjoint events. For example, the syntactical element *!pred(dk,_,_)* is used for declaring the disjointness of all facts *pred(X,Y,Z)* which share the same value of the attribute X. The attribute set defined by the disjointness clause is called disjointness key.

For an elaborated description of the complete syntax and semantics of probabilistic Datalog and the evaluation process refer to [Fuh95] and [FR95].

7.3 Modelling General Imaging using Probabilistic Datalog

Table 7.2 shows an example of computing the probability $P(d_1 \rightarrow q) = 0.533$. Assuming a document collection $D = d_1, ..., d_n$ we can derive a term vector for each document which represents the occurrence of a term within a document ($\tau(t_i, d_i)$). Given the probability distribution $\mu(t_i)$, a term t_i is regarded as a possible world. A document is regarded as a proposition which is true at a possible world, if the term occurs within the document. Column t_d assigns to each world where d_1 is not true the nearest worlds where d_1 is true. The probabilities of the non-d_1-worlds are transferred to the nearest d_1-worlds in column $\mu_{d_1}(t_i)$. For example, the probability of world t_2 is transferred to the worlds t_1 and t_6. The opinionated probability function $P^{t'}_{d_1}(t_i)$ transfers a portion of $\mu(t')$ to the world t_i. Considering the example given in Table 7.1 the nearest world gets 2/3 and the second nearest gets 1/3 of the probability. The resulting probability $P(d_1 \rightarrow q) = 0.533$ is computed as the sum of the (new) probabilities of the d_1-worlds where q is true.

As an example, consider the computation of $\mu_{d_1}(t_1)$:

$$\begin{aligned} \mu_{d_1}(t_1) &= \mu(t_1) * P^{t_1}_{d_1}(t_1) + \mu(t_2) * P^{t_2}_{d_1}(t_1) + \mu(t_4) * P^{t_4}_{d_1}(t_1) \\ &= 0.2 * 1.0 + 0.1 * 0.67 + 0.2 * 0.33 = 0.333 \end{aligned}$$

Now we are going to implement the above information on term probability, term occurrence and opinionated probability function in probabilistic Datalog.

This program defines the probabilities of the terms t_i (relation *term*) and the occurrence of the terms in a document (relation *docTerm*). The facts of *term* are declared to be disjoint (*!term(_)*). This corresponds to the disjointness

0.2 $term(t_1)$.
0.1 $term(t_2)$.
0.05 $term(t_3)$.
0.2 $term(t_4)$.
0.3 $term(t_5)$.
0.15 $term(t_6)$.
$!term(\_)$.
$docTerm(d_1,t_1)$.
$docTerm(d_1,t_5)$.
$docTerm(d_1,t_6)$.

0.67 $transfer(d_1,t_2,t_1)$.
0.33 $transfer(d_1,t_2,t_6)$.
0.67 $transfer(d_1,t_3,t_5)$.
0.33 $transfer(d_1,t_3,t_6)$.
0.67 $transfer(d_1,t_4,t_5)$.
0.33 $transfer(d_1,t_4,t_1)$.
$!transfer(dk,dk,\_)$.

about(D,T) :- docTerm(D,T) & term(T).
about(D,T) :- transfer(D,T',T) & term(T').

Figure 7.2: A probabilistic Datalog program

of possible worlds concerning the Imaging model. For the semantics of probabilistic Datalog, disjointness means that there exists no world where more than one fact of the relation *term* is true. The relation *transfer* is used to determine the portion of the probability of a term t_i which does not occur in a document to be transferred to a term which does occur. Thus this relation is used to express the shifting of the probabilities of non-y-worlds to the y-worlds modelling the opinionated probability function. The facts of *transfer* are disjoint with respect to the same disjointness key. The clause (*transfer(dk,dk,_)*) indicates that the first and second attribute form the disjointness key.

The query

?- about(d_1,t_1).

yields (0.333 ()) as answer, since

$$\begin{aligned}
\omega(about(d_1,t_1)) &= P(\eta(about(d_1,t_1))) \\
&= P(docTerm(d_1,t_1) \wedge term(t_1) \vee \\
&\quad transfer(d_1,t_2,t_1) \wedge term(t_2) \vee \\
&\quad transfer(d_1,t_4,t_1) \wedge term(t_4)) \\
&= 1.0 * 0.2 + 0.67 * 0.1 + 0.33 * 0.2 = 0.333
\end{aligned}$$

We formulate the whole query for all documents about t_1, t_4, and t_6 as

q(D) :- *about(D,t_1).*
q(D) :- *about(D,t_4).*
q(D) :- *about(D,t_6).*
?- *q(D).*

The result is (0.533 (d_1)).

7.4 Implementing General Imaging on top of Probabilistic Datalog

In order to implement General Imaging on top of Probabilistic Datalog we need to produce the necessary Probabilistic Datalog facts. To do so we need:

1. a "prior" probability distribution over the term space;
2. a measure of similarity between terms;
3. an opinionated probability function to direct and weight the transfer of probability between terms.

Once we have decided on these requirements, see for example in [Cv95] for an experimental setting, we need to set up one or more processes that construct the Probabilistic Datalog facts.

Figure 7.3 shows that in the first phase we may use information on inverse document frequency and term frequency to build the probabilistic relations *term* and *docTerm* as described in Figure 7.2.

How can we now determine the weights of the transfer facts which model the opinionated function? The portion is computed by the formula:

$$\omega(transfer(d, t_i, t_j)) = \frac{sim(t_i, t_j)}{\sum_{t|\tau(t,d)=1} sim(t_i, t)}$$

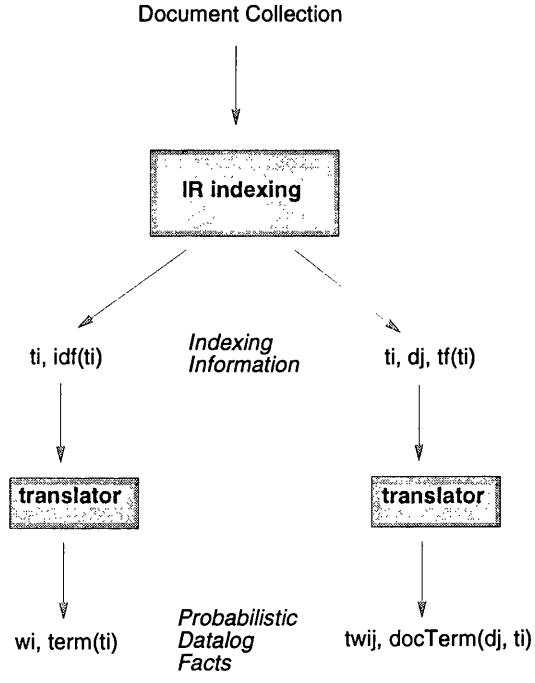

Figure 7.3: First phase of the construction of Probabilistic Datalog facts from IR indexing.

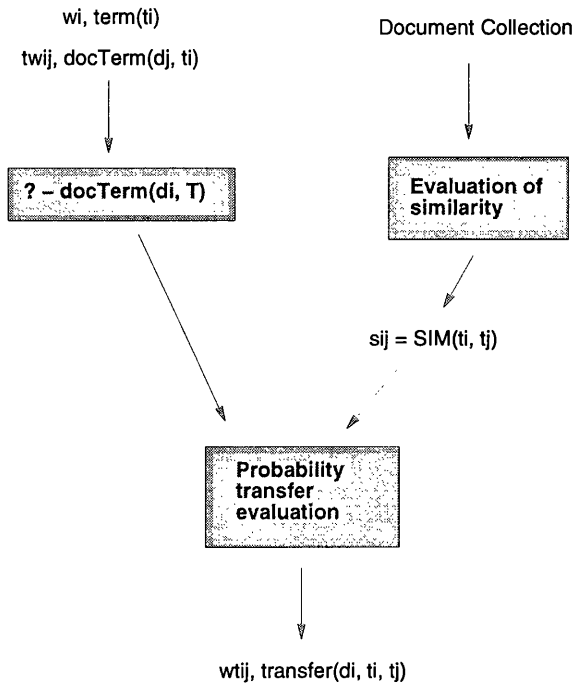

Figure 7.4: Second phase of the construction of Probabilistic Datalog facts from a index term similarity matrix.

The portion depends both on the similarity of the terms and on the set of terms occurring in the document. Figure 7.4 depicts the building of the transfer facts with the corresponding weights.

We use the result of phase 1 to determine the set of terms that occur in a document. Given a similarity measure we can then compute the weight of the transfer facts as defined by the above equation.

I have also developed rules to model probability transfer in the cases of incomplete and/or imprecise knowledge of the Term Space, a case that has been discussed in Chapter 6.

7.5 Conclusions

From the use of Probabilistic Datalog as an implementation platform for General Imaging we gain the possibility to combine the probability kinematics defined by General Imaging on terms with other probabilistic knowledge. The expressiveness of the knowledge representation and query language is

increased, because typical IR knowledge and queries may be combined with typical DB knowledge and queries.

The 2-phases approach presented above allows for implementing any kind of probability function, since the computation of the weight is done externally. A subset of the possible transfer functions could be computed using Probabilistic Datalog itself. But for the chosen transfer function I have not found a solution to compute it internally.

Chapter 8

Logical Imaging and Probabilistic Logic

In this chapter I explore an alternative representation of the Imaging revision methods, i.e. one that exploits a first-order logic for (objective) probability such as the $\mathcal{L}_1$ logic proposed by Halpern [Hal90]. I explore some of the consequences of this representation, as a possible implementation platform for RbLI and RbGLI.

8.1 Introduction

Imaging probability revision methods have originally been devised as mechanisms for giving semantics to conditional logic [Sta81]. Sebastiani [Seb96] argued that the application of Imaging in the context of the computation of relevance in IR is based on a somewhat non-standard interpretation of Imaging, as:

- the representation language is not that of propositional logic but a language of simple propositional letters, each representing a document or an information need;
- possible worlds are keywords; this means that there are not necessarily 2^n possible worlds, but there are as many possible worlds as the number of keywords in the application domain.

In this chapter I describe a representation of the RbLI model in terms of the $\mathcal{L}_1$ logic, as proposed first by Sebastiani in [Seb96] and later extended by Crestani, Sebastiani and Van Rijsbergen [CSvR96]. This representation of Imaging should solve the above problems and give a more standard interpretation of the Imaging process.

The chapter is structured as follows. Section 8.2 briefly described the $\mathcal{L}_1$ probabilistic logic. Section 8.3 describes how RbLI can be represented and implemented using the $\mathcal{L}_1$ logic. Section 8.4 concludes the chapter showing the advantages and the drawbacks of such approach.

8.2 The $\mathcal{L}_1$ probabilistic logic

The $\mathcal{L}_1$ probabilistic logic is a first order logic for reasoning about objective probabilities [BE90, Nut80, San89]. Probability values can explicitly be mentioned in the language: rather than mapping non-probabilistic formulae on the real interval $[0, 1]$, probabilistic formulae are mapped on the standard truth values *true* and *false*.

The logic allows the expression of real-valued terms of type $w_{\langle x_1,\ldots,x_n\rangle}(\alpha)$ (where α is a standard first order formula), with the meaning "the probability that random individuals $x_1, \ldots, x_n$ verify α". It also allows their comparison by means of standard numerical binary operators, resulting in formulae that can be composed by the standard sentential operators of first order logic.

The semantics of the logic is given by assuming the existence of a discrete probability structure on the domain; a term such as $w_{\langle x_1,\ldots,x_n\rangle}(\alpha) \geq r$ is true in an interpretation iff the probability assigned to the individuals that verify α sums up to at least r. It follows that, if x does not occur free in α, the term $w_{\langle x\rangle}(\alpha)$ may evaluate to 0 or 1 only, depending on whether α evaluates to *false* or *true*, respectively. Given a closed formula α, the term $w_{\langle x\rangle}(\alpha)$ plays then the role of its characteristic function.

The semantics of $\mathcal{L}_1$ can be specified by means of *type 1 probabilistic structures* (PS_1) for $\mathcal{L}_1$, i.e. triples $M = \langle D, \pi, \mu\rangle$, where:

- D is a domain of individuals;
- π is an assignment of n-ary relations on D to n-ary predicate symbols, and of n-ary functions on D to n-ary function symbols ($\langle D, \pi\rangle$ is then a first order interpretation);

- μ is a discrete probability distribution (DPD) on D.

The numerical value $\mu(d)$ may be interpreted as "the probability that, if a random individual has been picked from the domain D, individual d has been picked". In what follows, I will use $\mu(D')$ (where $D' \subseteq D$) as a shorthand for $\sum_{d \in D'} \mu(d)$. Also, given a DPD μ on D, μ^n is defined as the DPD on D^n such that $\mu^n(\langle d_1, \ldots, d_n \rangle) = \mu(d_1) \times \ldots \times \mu(d_n)$.

The $\mathcal{L}_1$ probabilistic logic is discussed in detail in [Hal90]. Rather then report here more detailed characteristic of this logic, in the following I will concentrate in describing the characterisation of the RbLI model using such logic.

8.3 Implementation of RbLI on top of probabilistic logic

In order to represent the RbLI model, a first subset of formulae is necessary to identify keywords and documents. This is necessary, as the domain of interpretation must be restricted to deal with these types of individuals only, which are the only entities of interest in the revision processes. Assuming that $T = \{t_1, \ldots, t_n\}$ is the set of terms by means of which documents are represented, and that $D = \{d_1, \ldots, d_m\}$ are the documents in our collection, we need the following formulae:

$$
\begin{array}{l}
Term(t_1) \wedge \ldots \wedge Term(t_n) \\
Document(d_1) \wedge \ldots \wedge Document(d_m) \\
\forall x.[x = t_1 \vee \ldots \vee x = t_n \vee x = d_1 \vee \ldots \vee x = d_m] \\
\forall x.\neg(Document(x) \wedge Term(x))
\end{array}
$$

This is a key feature of this approach as well as of the approaches of the implementation of imaging on top of Probabilistic Datalog: documents and terms are individuals belonging to the domain of discourse of a first order interpretation. In the ad hoc implementation presented in Chapters 3 and 4, instead, terms are (propositional) interpretations and documents are propositions.

The next subset of formulae is the one that specifies term occurrence, i.e. which documents are indexed by which term. We represent this by the for-

mula:

$$w_x(Occ(t_i, d_j)) = o_{ij} \qquad o_{ij} \in \{0, 1\}$$

for all $i = 1, \ldots, n$ and $j = 1, \ldots, m$, where o_{ij} is 1 iff t_i occurs in d_j.

Next, the probability of each term t_i is specified by means of the set of formulae

$$w_x(x = t_i \mid Term(x)) = p_{t_i} \qquad p_{t_i} \in [0, 1]$$

for all $i = 1, \ldots, n$. These formulae account for the case in which we want to input the probability values p_{t_i} from the outside. Alternatively, these probability values can be computed within $\mathcal{L}_1$ from the already available occurrence data, e.g. as their inverse document frequency (*idf*). In this case, the above formula is substituted by formula

$$w_x(x = t_i \mid Keyword(x)) = -log(w_y(Occ(t_i, y) \mid Document(y)))$$

The above formula compute the probabilities of keywords as their inverse document frequency; the formula $w_y(Occ(t_i, y) \mid Document(y))$ is in fact to be read as "the probability that, by picking a random document y, keyword t_i occurs in y". For the above to truly represent *idf*, though, we must assume that documents are picked with equal probability, which we state by formula

$$\forall xy.(Document(x) \wedge Document(y)) \Rightarrow [w_z(x = z) = w_z(y = z)]$$

which is to be read "if x and y are documents, the probability that by picking an individual at random x is picked is equal to the probability that by picking an individual at random y is picked". Alternatively, one may choose to include the previous three formulae in the representation. In this way, probability values are pre-computed "externally" and inputed to the reasoning process acting as "integrity constraints". This process is very similar to the one used in Chapter 7 dealing with the implementation of RbLI and RbGLI on top of Probabilistic Datalog. In what follows I will use the expression $P(t_i)$ as a shorthand of the expression $w_x(x = t_i \mid Term(x))$.

The next subset of formulae is the one that specifies the similarity between terms, i.e. the accessibility relation between worlds. A measure of how similar term t_i is to term t_j for all $1 \leq i, j \leq m$, $i \neq j$, is give by:.

$$Sim(t_i, t_j) = s_{i,j}$$

Only similarities between non equal terms are specified; in fact the case $i = j$ is not interesting for imaging methods, and its specification would complicate the formulation of following formulae. Values $s_{i,j}$ are input from an external source of information. Alternatively, they can be computed from within $\mathcal{L}_1$ from the already available occurrence values; for instance, they may be taken to be equivalent to the degree of coextensionality of the *Occ* predicate and computed by means of the formula:

$$Sim(t_i, t_j) = w_x(Occ(t_i, x) \mid Occ(t_j, x)) \cdot w_x(Occ(t_j, x) \mid Occ(t_i, x))$$

or else be computed according to some other measure of similarity, like for example the EMIM measure. On the other hand, the above formulae may act as integrity constraints. Further integrity constraints may be added if one's theory of similarity requires one to do so, in order to state further properties of similarity, like for example symmetry.

The following subset of formulae specifies, for each term, how the most similar term can be computed within $\mathcal{L}_1$ from the already available similarity data:

$$MostSim(t_i, t_{k_i}) \Leftrightarrow \neg\exists t_j.[Sim(t_i, t_j) \geq Sim(t_i, t_{k_i})]$$

Alternatively, one can input the "most-similar" values ($MostSim$) from the outside.

Next, we have to show how to calculate the revised probability of term t_i by imaging on document d_j, i.e. how to implement the probability transfer function. The revised probabilities are specified by the following numerical terms, for $1 \leq i \leq n$:

$$\begin{aligned} P_{d_j}(t_i) = w_x(Occ(t_i, d_j)) \cdot [P(t_i) + \\ + \textstyle\sum_{k=1}^{n} [P(t_k) \cdot w_x(\neg Occ(t_k, d_j)) \cdot w_x(MostSim(t_k, t_i))]] \end{aligned}$$

In order to compute $P_{d_j}(q)$ we now have to indicate by which terms the information need q is indexed:

$$w_x(Occ(t_i, q)) = o_i \qquad o_i \in \{0, 1\}$$

The probability $P_{d_j}(q)$ evaluated by imaging on d_j may be then calculated as:

$$P_{d_j}(q) = \sum_{i=1}^{n} w_x(Occ(t_i, q)) \cdot P_{d_j}(t_i)$$

The above modelling of the RbLI model can be easily extended to model RbGLI, by modifying the $MostSim$ and the $P_{d_j}(t_i)$ formula to account for the different kinematics of probabilities.

It is worthwhile to notice that, similarly to what happens in the implementation of imaging on top of Probabilistic Datalog [RÖ95, CR95], practically all the entities that participate in the imaging process are given here an explicit representation in the language of $\mathcal{L}_1$. However, unlike the implementation of imaging on top of Probabilistic Datalog, in the implementation of imaging on top of the $\mathcal{L}_1$ probabilistic logic an explicit representation is given even to:

- the formula that computes the prior probabilities of keywords;
- the formula that computes the similarities between keywords;
- the formula that chooses the recipients of a probability transfer and computes the revised probabilities of these recipients.

This hints to the fact that different formulae encoding different methods of computation of the above features may be experimented with [CSvR96]. In this sense, the whole information retrieval process is modelled as *a proper theory* of $\mathcal{L}_1$, whose role is that of a platform for experimentation of different models. Such a proper theory is obtained by assembling together various sets of formulae, each representing a class of entities participating in the process.

8.4 Conclusions

The above explanation of the implementation of RbLI on top of the $\mathcal{L}_1$ probabilistic logic suggests that $\mathcal{L}_1$ is a convenient and powerful platform for fast prototyping since it enables the evaluation of all the information necessary to the imaging process internally, as opposed to their external evaluation required by the implementation of imaging on top of Probabilistic Datalog.

As pointed out by Sebastiani in [Seb96], there are both advantages and disadvantages to having an internal definition/computation of the similarities between terms and their prior and posterior probabilities. The $\mathcal{L}_1$ approach has the advantage to be more self-contained and conceptually attractive, as it requires a minimum amount of data to be provided from outside the reasoning mechanism. Moreover, with a minimal coding effort, different probability kinematics methods may be experimented with and compared. In fact, a number of variants of the RbLI and RbGLI models have been presented in [CSvR96].

The price to be paid for this is that of efficiency, as reasoning in Probabilistic Datalog, a less expressive reasoning tool than $\mathcal{L}_1$, is no doubt more computationally tractable. Note that only "theoretical" tractability considerations are taken into account here. The $\mathcal{L}_1$ logic has not been implemented yet, while Probabilistic Datalog has, so an experimental comparison cannot be made yet. On this respect, it is plausible to think that data that needs to be computed once for all (such as similarity data between keywords) may be more efficiently computed outside the logic and subsequently fed to it. On another respect, the possibility to express the probability kinematics methods within the logic definitely seems desirable, at the very least if one conceives the logic as a fast prototyping tool.

Part V

Experimental Study

Chapter 9

The Troubles with Using a Logical Model of IR on a Large Collection of Documents

In this chapter I report on the challenges posed by trying to experiment with the RbLI model on large test collection of the size of TREC-B. The problems I found and the way I put together ideas and efforts to solve them are indicative of the troubles one might find in trying to implement and experiment with a "complex" logical model of IR. We believe our efforts could set an example for other researchers working on logical models of IR to try to implement their models in such a way that they can cope with the size of real life collections, though preserving the formal "beauty" of their logical models.

9.1 Introduction

In 1986 Van Rijsbergen [vR86] proposed the use of a non-classical conditional logic for IR. The proposal initiated a new line of research that was followed by many researchers (see for example [Nie88, Nie89, CC92, Bru93]).

A few years later Van Rijsbergen proposed to estimate the probability of the conditional by a process called Imaging [vR89]. This idea was finally put into an implementation in 1994 when Crestani and Van Rijsbergen [CvR95] (Chapter 3) proposed a retrieval technique called *Retrieval by Logical Imaging* (RbLI). This technique enables the evaluation of $P(d \rightarrow q)$ and $P(q \rightarrow d)$

by Imaging according to a "Possible Worlds" semantics where a term is considered as a possible world. This technique exploits term-term relationships in retrieval by means of an accessibility relation between worlds based on their Expected Mutual Information Measure (EMIM).

This chapter reports on the problems, solutions, and current results of the experimentation of Retrieval by Logical Imaging using a large collection of documents. The chapter is structured as follows. Section 9.2 lays out the experimental settings for the implementation of the model. This is where the problems start. Experimenting with a large collection, of the size of TREC-B, poses considerable difficulties that are reported in Section 9.3. Section 9.4 describes our attempted solutions towards an implementation of the RbLI model that could cope with the size of the test collection. Section 9.5 reports our current results in the context of the TREC-4 initiative, "ad hoc" track. Further directions of investigation are described in Section 9.6.

9.2 Implementing RbLI

As it has been introduced in previous chapters, in order to implement the RbLI model we require:

1. a "prior" probability distribution over the index term space that should reflect the importance of each index term in the term space;
2. a measure of similarity (or alternatively a distance) between index terms;

These two requirements reflects the use of a Possible World Semantics, since they correspond to the probability distribution, and to the accessibility relation measure among the possible worlds [Lew81].

The problem of determining an appropriate "prior" probability distribution over the set of terms used to index a document collection is one of the oldest problems in IR and many models have been proposed for this purpose. The problem could be translated into finding a so called "measure of the importance of the term in the term space". In IR several discrimination measures have been proposed (see for example [vR79, RS76]) and there it is not clear which one should be preferred to the others. For the experiments performed

in TREC I used the *Inverse Document Frequency* (*idf*) defined as:

$$idf(t) = -log\ \frac{n}{N}$$

where n is the number of documents in which the term t occurs, and N is the total number of documents in the collection.

Strictly speaking, this is not a probability measure since $\sum_t idf(t) \neq 1$, however since I assume it to be monotone to $P(t)$, we can use it instead of a proper probability function because we are only interested in a ranking of the documents of the collection, not in the exact probability values.

The problem of measuring the similarity between index terms in order to define a measure of accessibility among worlds is more difficult. It is very important to choose the most appropriate measure since much of the power of RbLI depends on it. For our TREC experiments I used the *Expected Mutual Information Measure* (EMIM). The EMIM between two index terms is often interpreted as a measure of the statistical information contained in the first term about the other one (or vice versa, it being a symmetric measure). EMIM is defined as follows:

$$I(i,j) = \sum_{t_i,t_j} P(t_i,t_j)\ log\frac{P(t_i,t_j)}{P(t_i)P(t_j)}$$

where i and j are binary variables representing terms.

When we apply this measure to binary variables we can estimate EMIM between two terms using the technique proposed in [vR77] p. 130. Using this measure we can evaluate for every term a ranking of all the other terms according to their decreasing level of similarity with it. We store this information in a file which is used at run-time to determine for an index term the most similar other index term that occurs in the document under consideration.

9.3 Experimenting with RbLI using a large document collection

In Chapter 3 the performance of the RbLI model was tested using a the *Cranfield* document collection. In Chapter 4 the RbLI model was generalised

into the RbGLI model and both these models were tested using the Cranfield, the *CACM*, and the *NPL* document collections. In the same chapter RbLI and RbGLI were compared with two prototypical classical retrieval models and I believe I have shown that *in principle* a probability transfer that takes into account a measure of similarity between the donor and the recipient is more effective in the context of IR than a probability transfer that does not take that into account.

In order to ensure that this result does not depend on the small size of the document collections I used, I decided to test these models on a collection of much larger size. Moreover, I decided to compare their performance with that of real IR systems, ones that could be recognised as a tough "benchmark".

I decided to proceed in two steps: first implement RbLI and test it, and only later implement RbGLI. The RbGLI model is more computationally demanding due to the introduction of a "probability transfer function" that enables the probability to be transferred not only to a single t_d but to a set of them according to their respective distance to the term under consideration.

The implementation and testing of RbLI in Chapters 3 and 4 [CvR95, Cv95] were quite heavy due to inefficient implementations of the probability transfer and to the complexity of the models. These problems, that were easily solved on small document collections, are much more difficult to tackle with a large document collection. In the following section I report on the challenge posed by trying to make the RbLI work on a large document collection. The Glasgow participation in TREC-4 was in the smaller part B collection which consists of 165,000 documents.

9.4 Getting RbLI to work

Because of the anticipated high computational load of performing RbLI on the TREC-B collection, it was decided to initially run experiments on a sub set of TREC-B so as to prototype the RbLI software being developed for these experiments. Rather than remove documents from the collection, it was decided to reduce all documents (which are in fact news articles) to just their lead paragraph, which generally for news articles is a summary of the article. This reduction resulted in a 70Mb document collection. All work reported in this section is based on this modified collection.

The implementations of RbLI reported in Chapters 3 and 4 were performed

on small test collections. Because of their size, it was possible to compute in a reasonable time the probability transfer of all terms in each document in the collection. On the TREC-B collection however, it was calculated that to perform this complete transfer would take too long given current computing resources. So methods of optimising the probability transfer were investigated.

9.4.1 Reducing the number of transfers

The first area looked at was the accessibility relation used to determine how probabilities are transferred, namely the EMIM measure. One of the features of EMIM is its ability to compute the relatedness between any two terms even if those terms don't co-occur, which means that in the case of RbLI it would be possible to compute probability transfers between all terms. The EMIM measure calculated between terms that don't co-occur however was found to be close to a small constant value, so for the experiments reported in this section a decision was made to restrict the EMIM calculations to only those terms that co-occur at least once. When transferring probabilities onto a document's terms, any term that doesn't co-occur with that document's terms will have its probabilities uniformly distributed to all those document terms, so as not to loose its probability.

By calculating EMIM between only co-occurring terms, the total number of calculations to be performed is reduced by around 95%. But it was felt that this reduction could and should be improved with further optimisations.

9.4.2 More speed

Now that probability transfers were reduced to just those terms that co-occur, the speed of RbLI on a document collection becomes proportional to the number of term co-occurrences in that collection. Therefore if we want to speed up RbLI, we need to reduce the number of these co-occurrences. There are a several ways in which this might be done. For example one could choose to only use those term co-occurrences where the two co-occurring terms appear in the same paragraph. Indeed this possibility might be exploited in the future, for the time being however it was decided to investigate the speed increase on RbLI when whole terms are removed from the collection. This technique was already proposed and tested in Chapter 3, where an intuitive explanation of its usefulness was given. I wanted to test its effectiveness on

a large document collection.

In choosing terms to be removed, the question arises which type of terms should we concentrate on: (a) the few terms that occur in many documents, or (b) the many terms that occur in a few documents, indeed is there any difference between the two? To answer this question, we need to examine the imaging process in more detail.

The part of RbLI that is the most computationally intensive is the final stage where probabilities are transferred onto the terms of each collection document. The time taken to complete this stage is proportional to the total number of probability transfers that will potentially be made. The term "potential" is used because not all transfers will happen. RbLI demands that even if a term co-occurs with several document terms, that term will only transfer its probability to just one of those document terms, its most similar. Nevertheless each one of these potential transfers has to be considered by the RbLI software, so each potential transfer does add to the time taken to complete this task. The number of potential transfers can be calculated using the following formula:

$$number\ of\ transfers = \sum_{d=1}^{D} \sum_{t=1}^{T_d} O_t$$

where: D is the set of all documents in the collection, T_d is the set of terms contained in document d, and O_t is the number of terms that co-occur with t.

The formula for O_t is as follows:

$$O_t = \sum_{d=1}^{D_t} (N_d - 1)$$

where: D_t is the set of documents containing term t, and N_d is the number of distinct terms in document d

So, for example, given the choice of, case A, removing 1 term that occurs in 100 documents or, case B, removing 50 terms that each occur in 2 documents, we can use these formula to calculate which of these choices will reduce the number of transfers the most. For example, if we assume that each document

contains 10 distinct terms, then the reduction in the number of transfers resulting from the two term removal cases is as follows[1]:

For the first case

$$number\ of\ tranfers(A) = 100 * 1 * (100 * 9) = 90,000$$

For the second case

$$number\ of\ tranfers(B) = 50 * (2 * *1 * (2 * 9)) = 1,800$$

From this, we can conclude that efforts should be concentrated on reducing the small number of terms that occur frequently in the collection.

9.4.3 Reducing the small number of terms that occur frequently in the collection

In initial experiments a standard stop list (taken from Van Rijsbergen [vR79] p. 18) was used when indexing TREC, but in the light of the work described above, it was decided to investigate how retrieval performance would be affected when a bigger stop list was used.

Using a standard $tf - idf$ retrieval system, the effect on retrieval performance from using a number of different stop lists was tested. The definition of a term's membership for these stop lists was based on the number of documents that term occurred in. Stop lists were generated for terms that occurred in more than 90% of documents, more than 80% of documents, more than 70% of documents, and so on down to 2.5% of documents. For each stop list, the modified TREC-B collection was re-indexed, a retrieval run was performed and recall precision figures were obtained for each run. In addition two extra runs where performed where no stop list was used and a standard stop list from Van Rijsbergen [vR79] was used. The graph in figure 9.1 shows a selection of these runs.

As can be seem from the graph, precision is improved at all recall levels when a standard stop list is used instead of no stop list. If a stop list containing

[1]The removal of term(s) has another minor influence on the time taken to complete RbLI: all terms co-occurring with the term(s) being removed will have fewer probability transfers to them. The effect of this influence however is the same for both cases, and so it need not be considered.

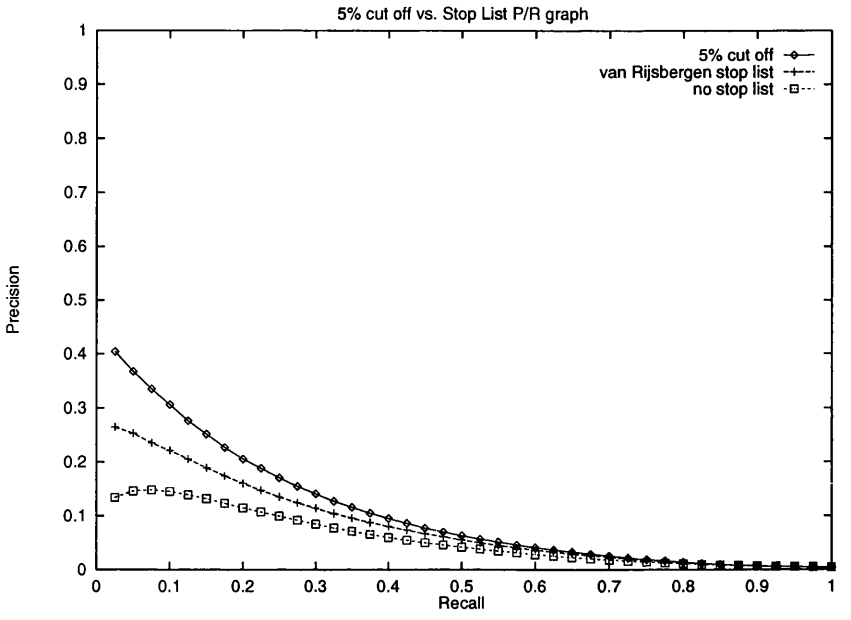

Figure 9.1: Precision and recall figures with different stop lists.

Initial number of occurrences	12,594,371
Occurrences after 5% stop list	6,902,676
Number of words in 5% stop list	253

Table 9.1: Effects of a 5% stop list

terms occurring in more than 5% of the documents is used however, there is a further uniform improvement in precision. The terms in this stop list account for around 50% of the term occurrences in the collection, as can be seen from Table 9.1. Although not plotted in this graph, it was found that using a larger stop list reduced precision.

From this experiment, it was concluded that the 5% stop list should be used when indexing the TREC-B collection so as to improve the speed of the RbLI process.

9.5 Evaluating Retrieval by Logical Imaging using the TREC-B document collection

Using the results reported in the previous section, I tried to perform a few experiments using the improved RbLI software on the *TREC-B document collection* for the *ad hoc* track. However, I soon run into a lot of small technical problems, the kinds of problems that almost all first time TREC participants experienced. My lack of experience in dealing with large document collections together with the complexity of the RbLI model, made it impossible to have retrieval results ready for the TREC deadline.

In August 1995, when it became clear that the RbLI experiments would not be ready in time for the TREC-4 deadline, we of the IR Glasgow group decided to use a more classical IR system we already had developed in Glasgow to produce the retrieval results for the "ad hoc" track to send to TREC-4. We thought that we could later use the results of this system as a benchmark for the RbLI results, when these would be ready. The *glair* result set was then quickly generated and submitted instead of the RbLI results, that were not yet available.

The benchmark system we adopted for comparison with RbLI is a "text book" IR system. It is based on the classical $tf - idf$ retrieval strategy. Terms found in documents and queries have first their case normalised, then, any of these terms appearing in a stop list were removed. The stop list was taken from [vR79], since the stop list experiments reported in the previous section had not been carried out at this stage. The remaining terms were suffix stripped using the Porter stemmer [Por80]. Document terms were weighted using the $tf - idf$ weighting scheme. The *idf* formula has already been defined in Section 9.2, tf is defined as in [FBY92]:

$$tf_{i,j} = \frac{\log(freq_{i,j} + 1)}{\log(length_j)}$$

where $freq_{i,j}$ is the frequency of term t_i in document d_j, and $length_j$ is the number of unique terms in document d_j.

The $tf - idf$ retrieval strategy simply evaluates the product of the two components and ranks the documents in the collection based on a score. The score for each document is calculated by summing the $tf - idf$ weights of any query terms found in that document.

We submitted to the TREC-4 Conference (TREC-B) the results of this system with low expectations. However, the results we achieved were not bad at all, as is summarised in the the official TREC-4 results [CRSvR95]. The system, in the context of the TREC-B only participants, gave the best performances in 10 queries, and the worst performances in 6 queries out of the 49 used in TREC-B. Its overall performance was well above the median value of the average precision.

The results of the $tf - idf$ retrieval strategy will be compared with the results obtained by RbLI and RbGLI in Chapter 10.

9.6 Conclusions

We have been told by others that there is a tradition that "TREC first timers" fail to get their planned experiments done by the required deadline. We unfortunately have done nothing to change this.

Trying to implement RbLI on the TREC collection proved to be a compromise between the theoretical purity demanded by the model, and the implementation problems posed by a collection of the size of TREC-B. We have found this compromise to a be a driving force in revealing other areas of work to be investigated. Therefore, experimental results aside, we regard our first participation in TREC as having been beneficial.

Chapter 10

Implementation, Experimentation and Evaluation Using a Large Collection of Documents

This chapter reports in detail on the implementation, experimentation and evaluation of the models proposed in the theoretical study (Chapters 3 and 4). The experimentation reported here makes use of a collection of documents much larger in size than those used in previous chapters.

10.1 Motivations

Chapters 3 and 4 reported the description and the results of a set of experiments carried out on various standard test collections of relatively small size. These tests were done mainly for checking, in a quick and easy way, the actual effectiveness of some of the theoretical ideas proposed. Although in the past it was perfectly acceptable to test new models on collections of such a small size, and in fact most of the early models of IR were tested in such way [Sv76], in recent years it has become almost necessary to test new models on collections of much larger size. As pointed out in Chapter 1, real life applications nowadays usually have to deal with hundreds of thousands of documents. In order to be sure that the performance of a model scales up to the requirements of present day applications, the model needs to be

tested on some large collection of documents.

In Chapter 9 we reported an attempt at experimenting RbLI using a larger test collection. This experimental evaluation was done in the context of the TREC-4 framework [Har95a], with fixed and strict deadlines. Unfortunately, the short time and the limited resources available at that time compared with the size of the task, did not allow us to complete the experimentation in time for the TREC-4 deadline. Later on, however, with more time and with the availability of a more powerful computer and larger disk space, this experimentation could be continued and completed. The experience of the TREC-4 attempt proved very useful.

This chapter reports on the the implementation, experimentation, evaluation, and failure analysis of the two models: Retrieval by Logical Imaging (RbLI) and Retrieval by General Logical Imaging (RbGLI) already presented in Chapters 3 and 4. This experimental investigation was carried out using a large collection of documents and an experimental IR system that will be briefly described in the next two sections.

10.2 The Wall Street Journal document collection

Until a few years ago, there was no large test collection available for performance testing in IR. Most of the evaluation of experimental IR systems was done using relatively small test collections. These collections were built with considerable efforts by the same people who were building experimental IR systems, since using them was the only way to prove the effectiveness of some of the theoretical ideas proposed. Examples of these collections are the CACM, the Cranfield, the NPL, the LISA, etc. The characteristics of some of these collections are reported in Chapter 4 (Table 4.5). A complete survey of all the collections used in the early days of IR is reported in [Sv76].

In recent years some developers of commercial IR systems pointed out the gap that was opening between the test collections in use for experimental purposes and the collections of documents used by commercial IR systems. The collections commercial IR systems had to manage were becoming several orders of magnitude larger than the test collections. Some researchers also started to question the validity of some of the effectiveness results obtained using test collections. The problem of the *scalability* of the techniques proposed and experimented using test collection became an important issue in

the IR research. The question that many researchers started asking was: are techniques proved to be effective for small test collections also effective for much larger collections?

Although a number of researchers started putting increasing effort into building larger test collections since the 80s, the first milestone in this direction was the starting in 1990 of the Defence Advanced Research Projects Agency (DARPA) *TIPSTER* project at the University of Massachusetts [Cro92], in the USA. However, what capture the attention of a large number of IR researchers was in 1992 a "spin off" of TIPSTER project, called TREC.

TREC (Text REtrieval Conference) is a workshop series sponsored by the National Institute of Standards and Technology (NIST) and DARPA that promotes large scale IR research by providing appropriate test collections, uniform scoring procedures, and a forum for organisations interested in comparing their results. The annual TREC is an event in which organisations with an interest in IR and information filtering take part in a coordinated series of experiments using the same experimental data. The results of these individual experiments are then presented at a workshop where tentative comparisons are made. In order to preserve the desired, pre-competitive nature of these conferences, the organisers have developed a set of guidelines constraining the dissemination and publication of TREC evaluation results. These guidelines are meant to preclude the publication of incomplete or inaccurate information that could damage the reputation of the conference or its participants and could discourage participation in future conferences[1].

The first TREC (TREC-1) was held in November 1992 at the NIST headquarter and saw the participation of 25 international groups [Har93]. Since then the event has assumed more and more importance in the IR research, so much that more than 75 groups showed interest in participating in the latest TREC (TREC-6) event. A detailed description of TREC and its history is outside the scope of this section. The interested reader can find further information about TREC in the annual proceedings of the TREC Conference held every year at NIST and in particular in the annual report prepared by Donna Harman [Har93, Har95b, Har94, Har95a, Har96]. Some critical reflections about the usefulness of the TREC initiative can instead be found in [SJ95].

Critical to the success of TREC was the creation of a set of tools for testing experimental IR systems. In particular, TREC has built and made available

[1]More details about TREC can be found on the TREC home page at NIST: http://www-nlpir.nist.gov/trec.

to IR researchers over the years a set of large test collections. These collections were built thanks to the contribution of publishers, news providers, and US government offices. For the purpose of the experimentation reported in this thesis, a subset of the main TREC collection was used. This subset is made up of full text articles published in the *Wall Street Journal* (WSJ) in 1990-92. This collection of documents, about 250 MB in size and accompanied by a set of 300 queries and relevance judgements, constitutes by all standards a large test collection. The main characteristics of this collection are reported in Table 10.1. A quick comparison with the collections used in Chapters 3 and 4 shows that the WSJ collection is at least one order of magnitude larger than the largest of them.

Data sets:	*WSJ-full*	*WSJ-lead*
num. of documents	74.520	74.520
size in MB	247	72
num. of queries	300	300
unique terms in documents	123.852	61.079
unique terms in queries	3.504	3.504
avg. document length	180	60
avg. query length	40	40
avg. num. of rel. doc.	35	35

Table 10.1: The Wall Street Journal document collection.

Table 10.1 reports two collections, the *WSJ-full* and the *WSJ-lead* obtained from the WSJ collection. The difference between these two collections is that, while WSJ-full is the actual WSJ collection from TREC, composed of the full text of the documents, WSJ-lead is only composed on the leading paragraph of the documents. WSJ-lead is therefore not an official TREC collection and was built on purpose for this thesis.

The reason for the use of these two collections lies in the difficulties found in experimenting the proposed models using a large term space. The use of only the leading paragraphs of the documents reduces the term space to about one half of its original size, while still enabling the use of the 300 queries and full relevance judgements accompanying the collection. This reduction of the terms space has even more influence on the size of the EMIM data and on the time requested to calculate them (see Chapter 9). The identification of the leading paragraph of a document is a simple task, since documents are marked up using SGML. An example of a document of the WSJ collection is reported in the following. The leading paragraph is marked by the $< LP >$

tag.

```
<DOC>
<DOCNO>
WSJ900402-0192
</DOCNO>
<DOCID>
900402-0192.
</DOCID>
<HL>
   VLSI to Post Profit
   Matching Forecasts
   For the First Quarter
</HL>
<DATE>
04/02/90
</DATE>
<SO>
WALL STREET JOURNAL (J), PAGE A8B
</SO>
<CO>
VLSI
</CO>
<IN>
DOW JONES INTERVIEW (CEO)
</IN>
<LP>
   NEW YORK -- VLSI Technology Inc.'s first-quarter earnings
should meet analysts' expectations, the company's chairman
and chief executive officer, Alfred J. Stein, said.
   "We expect to do as well as the analysts are projecting .
. . between five and eight cents a share," Mr. Stein added.
VLSI makes standard and customized integrated circuits.
</LP>
<TEXT>
   Mr. Stein noted that late last year, the company guided
analysts' first-quarter projections lower from earlier
estimates of around 15 cents a share.
   He said a slowdown in standard chip-set sales and a drop
in demand for custom chips by its largest customer, Apple
Computer Corp., stalled revenue growth in the quarter.
   The company had a loss of $6.3 million in the first
quarter of 1989, largely due to problems during the start-up
of its chip plant in San Antonio, Texas. VLSI earned 11 cents
a share in the latest fourth quarter.
   The company expects to release its first-quarter earnings
April 12.
   Mr. Stein said seasonal slowdown in Far Eastern demand for
```

```
the chip sets was partly responsible for damping growth in
the first quarter.
   The region's IBM-compatible computer makers use the sets
in personal computers that see their strongest sales before
the Christmas holidays. Far Eastern demand generally slacks
off in the first quarter, Mr. Stein said. Shipments to the
Far East account for about half of the company's chip-set
sales.
   Mr. Stein said demand for customized chips by Apple
Computer has recovered from a drop that also depressed
first-quarter revenue.
   "Apple is coming back very strongly to us," Mr. Stein
asserted. He added, however, that the impact of Apple's
resumed demand won't be felt until the second and third
quarters of this year.
   Apple accounted for 13% of VLSI's revenue in 1989, while
sales to International Business Machines Corp. rose to about
10% of total revenue.
   Mr. Stein said sales to IBM will exceed sales to Apple
this year because of increasing shipments to IBM, not because
of shrinking sales to Apple. The increasing importance of IBM
as a customer illustrates VLSI's strategic shift toward sales
of "application-specific standard product," primarily
standardized chip sets for personal computers, over the
customized chips designed for Apple Computer and others.
</TEXT>
</DOC>
```

An example of one of the 300 queries (called topics in TREC) accompaining the WSJ collections is reported in the following:

```
<top>
<head> Tipster Topic Description
<num> Number: 003
<dom> Domain: International Economics
<title> Topic: Joint Ventures
<desc> Description:
Document will announce a new joint venture involving a Japanese
company.
<narr> Narrative:
A relevant document will announce a new joint venture and will
identify the partners (one of which must be Japanese) by name, as well
as the name and activity of the new company.
<con> Concept(s):
1.  joint venture, tie up
```

```
2.  partner, cooperation, joint management, agreement
3.  cooperate, work together, jointly manage, jointly own, jointly
produce
<fac> Factor(s):
<nat> Nationality: Japan
<time> Time: Current
</fac>
<def> Definition(s):
Joint Venture - An international business undertaking defined as the
commitment, for more than a very short duration, of funds, facilities,
and services by two or more legally separate interests to an
enterprise for their mutual benefit.
</top>
```

10.3 The SIRE experimental IR system

SIRE (System for Information Retrieval Experimentation) is a prototype indexing and retrieval toolkit developed by Mark Sanderson at the Department of Computing Science of the University of Glasgow. SIRE is a collection of small independent modules, each conducting one part of the indexing, retrieval and evaluation tasks required for classic retrieval experimentation. The modules are linked in a pipeline architecture communicating through a common token based language. SIRE was initially used in research examining the relationship between word sense ambiguity, disambiguation, and retrieval effectiveness [San96b]. It proved to be a flexible tool as it not only provided retrieval functionality, but a number of its core modules were used to build a word sense disambiguator as well. It was also used in the experiments for the Glasgow IR group submissions to TREC-4 (reported in Chapter 9 and in [CRSvR95]), TREC-5 [SR96], and TREC-6 [CSTL97] and is currently being used in a number of research efforts within the group. The system has also been successfully used by many students of the Master of Science in Advance Information Systems of the University of Glasgow for their practical work.

SIRE is implemented on the UNIX operating system which, with its scripting and pre-emptive multi-tasking is eminently suitable for handling the modular nature of SIRE.

A detailed description of the functionalities of SIRE is outside the scope of this section. The system is currently available on public domain for research purposes. The interested reader should contact Mark Sanderson for a copy of a short unpublished paper describing the system [San96a] and for the

location of SIRE's binary files.

10.4 Implementation of RbLI and RbGLI on top of SIRE

The flexibility of the SIRE experimental system, the availability of its source code, and the willingness to help of SIRE's creator (Mark Sanderson), made SIRE the most obviuos choice as an implementation platform for experimenting RbLI and RbGLI with a large collection of documents. To this choice, three other factors contributed:

- The code developed for the experiments reported in Chapters 3 and 4 was highly inefficient and unsuitable for experimenting on a large scale. This was developed simply for testing the theoretical ideas reported in those chapters in the fastest possible way. A re-ingeneering of that code would have been too time consuming.
- The current implementation of Hyspirit, the system developed at the University of Dortmund implementing Probabilitic Datalog [FR97], was not efficient enough to enable experimenting on a large scale. Therefore, despite the ease of implementing RbLI and RbGLI on top of Probabilistic Datalog (see Chapter 7), it was decided that that was not a suitable implementation platform, since it would not have enabled an experimentation using a large collection of documents.
- The $\mathcal{L}_1$ logic (Chapter 8) has not been implemented yet.
- SIRE had been successfully used for the Glasgow partecipation in TREC-4, TREC-5, and TREC-6 with collections even larger than WSJ. In these tasks it proved to be highly flexible and at an acceptable level of efficiency.

Once SIRE was chosen as the implementation platform for RbLI and RbGLI, it was necessary to develope a number of new modules for the experimentation carried out in this thesis. These new modules were build with the same design principles of the already existing modules, so that they could be used in a pipeline with them. Some of these modules were written using the C programming language [KD88], some others using the Perl scripting language [WCS96].

In the following, no detailed description of the implementation of RbLI and RbGLI using SIRE will be reported, since this is not of much scientific interest. Instead, the rest of this chapter will concentrate on reporting and analysing the results of the experimentation.

10.5 Experiments with the RbLI model

The first set of experiments reported in the following concerns the implementation and evaluation of the RbLI model. This is a continuation of the effort started in the framework of TREC-4 (see Chapter 9). In that context, considerable effort went into finding effective ways to evaluate the accessibility relation by means of EMIM and to cut down the number of probability transfers. Those results have been used for the experiments reported in this chapter.

The weight assigned to terms in the term space is the *idf* weight, calculated as:

$$idf_i = log(\frac{N}{n_i})$$

where n_i is the number of documents in which the term t_i occurs, and N is the total number of documents in the collection.

Although the *idf* weight cannot be considered a probability, nevertheless we can assume it to be an approximation of the probability. We can use this estimate because we are not really interested in finding exact probabilities, but only to produce a ranking of the documents according to them. In the following the term "weight" and "probability" will be used interchangeably.

Regarding the experimental design, for the purpose of the experimentation of this model, and as a matter of fact for the entire experimentation, it was decided to follow a "incremental" approach. Experiments were first performed on the "easiest" model (RbLI) and on the smallest document collection (WSJ-lead). This combination enabled to perform a large number of experiments in a short time. Once results were acquired and the best combination of parameters was achieved, this setting would be used to perform experiments on a larger document collection (WSJ-full) and on a more complex model (RbGLI). Such approach enabled to achieve the best combination of parameters in the fastest time. Because of that, this section is

divided in two parts, the first part reporting the experiments carried out using the WSJ-lead collection and the second part reporting the experiments carried out using the WSJ-full collection.

10.5.1 Using only leading paragraphs

This section reports on a series of experiments carried out using the WSJ-lead document collection. The experiments here reported aimed at finding the best combination of parameters to achieve the highest level of effectiveness from the RbLI model in terms of classical IR measures of performance, that is recall and precision. Results are graphed using the 10 recall points evaluation procedure reported in [vR79].

Redistributing the untransfered probabilities

The first theoretical hypothesis to be tested is related to the case of incomplete knowledge of the accessibility relation. This case was discussed in Chapter 6.

In the experiments reported in this thesis, the accessibility relation was estimated using the Expected Mutual Information Measure (EMIM). This measure has been studied in details and used by many researchers in the past and an effective way of estimating it was proposed by van Rijsbergen in [vR79]. Nevertheless, EMIM is a very computational expensive measure to evaluate in the presence of a large term space. An evaluation of the full EMIM for all term pairs in the WSJ-lead collection would have required about 70 hours of CPU time using the machine available for these experiments, that is a Sun Ultra 2 running Solaris 2.5.1 with 256 MB of RAM. The storing of the EMIM data would have required about 24 GB of disk space.

In order to reduce the time and space required to evaluate EMIM we adopted the series of simplifications reported in Chapter 9, that made it possible to evaluate EMIM in just about 6 hours of CPU time and using only about 2.5 GB of disk space. However, this means that we had to deal with incomplete knowledge of the accessibility relation between terms in the term space.

From a theoretical point of view, the incomplete knowledge of the accessibility relation in the terms space creates the series of problems already discussed in Chapter 6. The solution to such problems lies in a redistribution of the untransfered probabilities over the terms occurring in the document (also

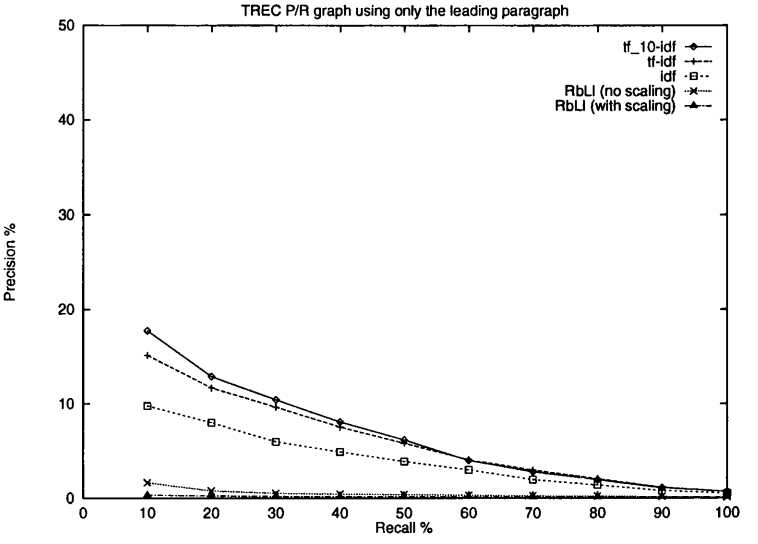

Figure 10.1: Precision and recall graphs for the WSJ-lead collection using the RbLI model with or without probability scaling.

called d-terms). This is equivalent to a *scaling* of the posterior (imaged) probabilities assigned to terms in the document to make their sum equal to one (i.e. making the document certain). This scaling is achieved using the same kinematics of the RbCP model.

Figure 10.1 reports the results of the use of the RbLI model with scaling and without scaling. The figure also reports, as a reference, the results using the classical *idf* and $tf - idf$ weighting schemas without probability transfers [Har92a]. The formula for *idf* has already been described, while tf is calculated as follows:

$$tf_{ij} = \frac{log(freq_{ij} + 1)}{log(length_j)}$$

where $freq_{ij}$ is the frequency of occurrence of term t_i in document d_j, and $length_j$ is the number of unique terms in document d_j.

The $tf_{10} - idf$ weighting schema is instead a modification of the classic tf-idf schema, proposed by Sanderson in [SR96], that can be obtained using the following formula for the evaluation of tf_{ij}:

$$tf_{10\ ij} = \frac{log((freq_{ij} \cdot 10) + 1)}{log(length_j)}$$

This weighting schema has proved experimentally to be slightly more effective than the classical $tf-idf$ schema when dealing with long documents [San96b].

It can be easily seen that the RbLI model is behaving very poorly, and that scaling of probabilities to make up for the incomplete knowledge of the accessibility relation makes it behave even more poorly. It was therefore decided not to perform the scaling for the succeeding experiments, since it only increases the number of computations and does not improve the performance of RbLI.

Notice that the low level of performance of $tf - idf$ and $tf_{10} - idf$ is due to the use of only the leading paragraph as document text. The levels of performance of these models increase considerably when the full text of documents is considered, as it will be shown further on in this chapter.

Using different amounts of EMIM data

A previous study carried out using small test collections showed that it is possible to reduce the size of the EMIM data to be stored and scanned at run time to direct the probability transfers, without any significant decrease in performance for the RbLI model (see Chapter 3). Another set of experiments was therefore directed towards using different amounts of EMIM data to see if it was possible to achieve this effect also on a large collection.

Figure 10.2 shows the performance of RbLI using different amount of EMIM data. The performance of RbLI using the full EMIM data (but with the simplifications described in Chapter 9) were compared with those achieved by RbLI using only a percentage of the full EMIM data.

With regards to the kinematics of probabilities, the use of $x\%$ of EMIM data is equivalent to the use of an accessibility relation that only captures the $x\%$ most similar terms to a given term and that does not provide any accessibility relation to the remaining (100 - x)% terms in the term space.

The percentages used here are approximative, since they were obtained by putting a threshold on the EMIM value between pairs of terms. The three values of 10%, 20%, and 30% of percentage of EMIM data are reported in the figure. Percentages over 30% did not show any difference in performance from the use of the full EMIM data.

Surprisingly, the performance of RbLI improves with the use of smaller amounts of EMIM data. The reason for this improvement in performance

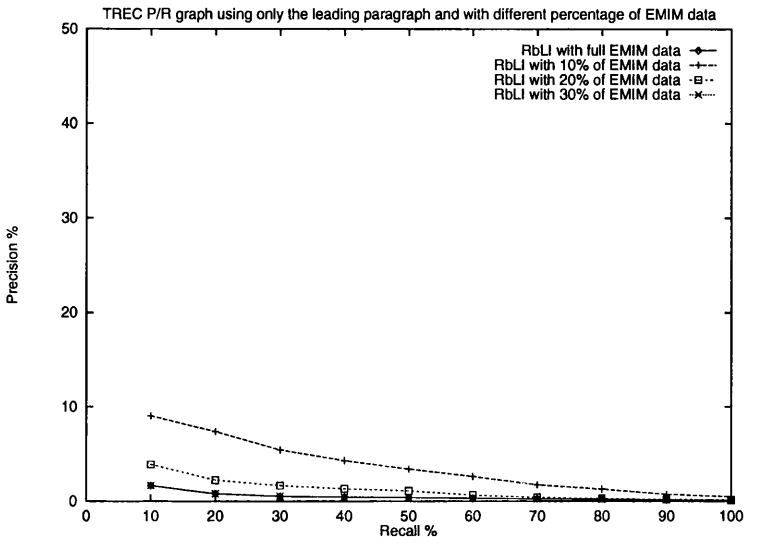

Figure 10.2: Precision and recall graphs for the WSJ-lead collection using different percentage of the full EMIM data.

can be found in the modification of the kinematics of the RbLI model that this causes. Probability is transfered from a not-d-term to a d-term only if there is a significant level of similarity between the two. By using the full EMIM data we can always find a d-term that is similar to a not-d-term, to which the probability of the not-d-term will be transfered, even if the similarity between the two is very low. The results reported in Figure 10.2 suggest that this is not an effective kinematics. Moreover, results reported in Figure 10.1 suggest that it is better not transfer at all the untransfered probability remaining attached to not-d-terms, than transfer it without good reasons. In other words, it is better to loose some probability (and therefore having a probability of the document less than one), than transfer it in an unjustified way.

Combining the above result with the result of the previous experiments, it seems that the combination of the kinematics of RbLI and RbCP is not an effective combination in the context of IR. This results disproves some of the theoretical ideas reported in Chapter 6. The effectiveness of the combination between RbGLI and RbCP will be tested later on.

Using different stoplists

Another set of experiments was devoted to find the best possible stoplist to be used with the RbLI model. There are a few reasons why it is important to study the term discrimination power and identify which terms could be removed from the term space. In fact, removing unnecessary terms from the term space:

- reduces the number of probability transfers that are going to be performed for each document in the collection;
- reduces the effort of evaluating the EMIM for the terms in the term space;
- enables to avoid to take into consideration terms that would give a very little contribution to the overall posterior probability of a document.

The identification of the terms to be removed from the term space should be related to the their discrimination power. Using the classical Zipfian distribution studied by Luhn [Luh57] in the context of IR, the terms that have the lowest discrimination power are those that are the most and the least frequent ones in the collection. In IR usually only the most frequent ones are removed, since removing the least frequent ones does not have many advantages. This is the direction followed in these experiments.

Moreover, in addition to the classical motivations for removing very frequent terms, another important motivation was discovered by looking at the distribution of the weights in the term space before and after imaging.

Figure 10.3 depicts for every term in the WSJ-lead term space the *idf*, that is the term weight before imaging (the prior probability), and the average weight of the term after imaging (the posterior probability). The graph clearly shows that frequent terms, i. e. terms with low *idf* weight, have a very high weight after the imaging process. This is an undesired effect of the way imaging is implemented by the RbLI and the RbGLI models. Terms that are very frequent in the collection will also co-occur very frequently with other terms. Since co-occurrence is the most important factor in the evaluation of the EMIM values (see Chapters 3 and 9), then two terms that co-occur often, have also a high EMIM value. This means that a very frequent term occurring in the document, i.e. a d-term with low *idf*, attracts a very large amount of weights from non-d-terms. Other d-terms receive very little

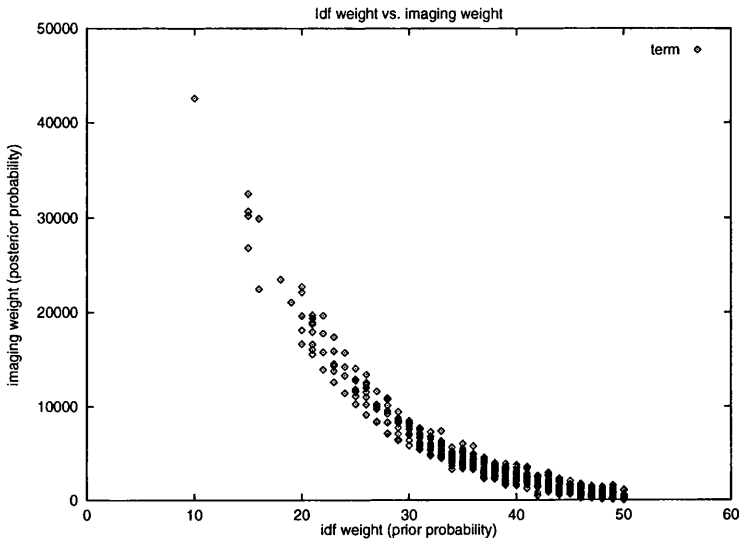

Figure 10.3: Prior probability (idf) vs. posterior probability (imaging weight).

weight from non-d-terms, since given a non-d-term the very frequent d-term will almost always be the one with the highest values of EMIM among the d-terms. This causes an unbalanced kinematics of the weights that may have disastrous effects in the retrieval phase.

In fact, let us suppose that t_i is a very frequent term, if t_i is a d-term for document d_j than it will attract a large weight since it will be for many non-d-terms the d-term with the highest EMIM value. If t_i is also a q-term (a term occurring in the query) then document d_j will be one of the document ranked at the top for query q. A document d_k for which t_i is not a d-term will probably be ranked at at lower position that d_j, even if d_k had a larger number of q-term than d_j.

Figure 10.4 reports the performance of RbLI with different sizes of the stoplist. The reference size is the standard stoplist reported in [vR79] comprising 320 stopterms. Other stoplists were built by adding to the standard stoplist the x% most frequent terms in the collection excluding those already present in the standard stoplist; for example, the so called "top 1%" stoplist contains 320+610 stopterms, where 610 is 1% of the total number of terms in the term space.

The results show that a 1% stoplist gives better performance than the standard stoplist. This is achieved by removing from the term space some of the terms that attract large weights during the imaging process. Stoplists larger

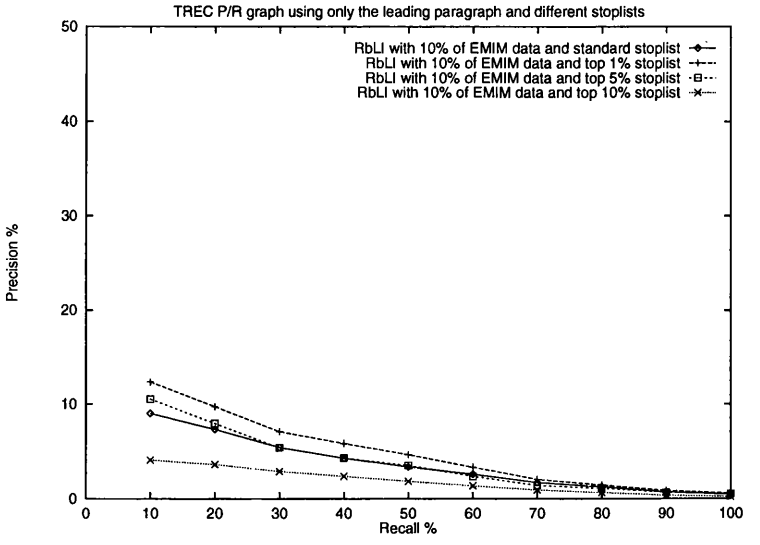

Figure 10.4: Precision and recall graphs for the WSJ-lead collection using different stoplists.

than the top 1% one, like the top 5% and the top 10% ones, remove also some terms that are useful for the retrieval process, like terms that are present in queries, therefore decreasing the performance levels.

Choosing the best combination of parameters

Using the results of the various experiments reported above, it was possible to devise the best combination of parameters for the RbLI model and the WSJ-lead collection. The results of a performance comparison between the RbLI model and the classical idf, $tf-idf$, and $tf_{10}-idf$ models are reported in Figure 10.5.

The results show that RbLI performs slightly better than the idf model (which is equivalent to the RbJP model of Chapter 4), although worse than the $tf-idf$, and the $tf_{10}-idf$ models. However, we need to remind here that only the idf model allows a fair comparison with the RbLI model, since the RbLI model does not use any information about the term distribution inside a document, like for example the tf weight, that is used by the $tf-idf$, and $tf_{10}-idf$ models.

The above result, although of a smaller magnitude, is in complete agreement with the result presented in Chapter 4, obtained on a much smaller term space.

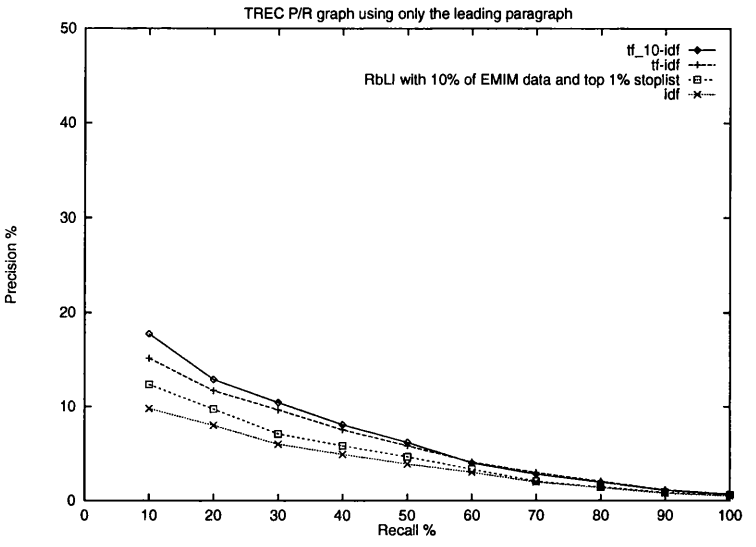

Figure 10.5: Performance of the RbLI model using the WSJ-lead collection.

10.5.2 Using full documents

This section reports on the experimental results obtained by using the WSJ-full collection. Given the size of the collection and the complexity of experimenting RbLI on such a large term space, only a limited number of experiments could be performed. The results of the experimentation with the WSJ-lead collection, presented in the previous section, were used to best direct the experimentation reported here.

The best combination of parameters

A large set of experiments, whose results I am not going to report here, suggest that the best combination of parameters (regarding scaling, amount of EMIM data, length of the stoplist, etc.) for RbLI on WSJ-full is the same that proved to be the best for RbLI on WSJ-lead.

Figure 10.6 show that the performance of the RbLI and *idf* models are very close, with RbLI performing slightly better at some recall levels.

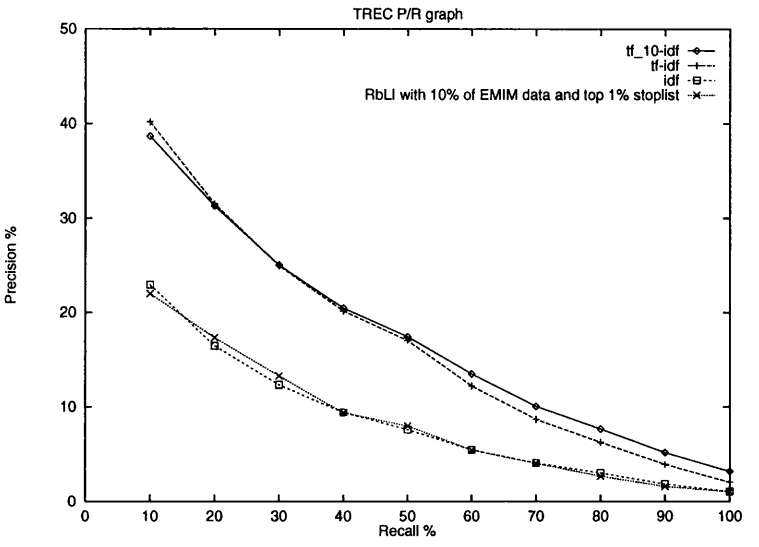

Figure 10.6: Precision and recall graph for the WSJ-full collection.

The document length effect

A recognised problem with all the models presented in Chapter 4, and therefore also with the RbLI model as it has been implemented in this section, is related to the document length. The problem can be easily explained looking at Figure 10.7.

Let us suppose we have two documents d_1 and d_2. The document d_1 is considerably larger in number of terms than d_2. Here we only consider the number of unique terms, so in actual facts d_1 and d_2 could be of similar size, but d_1 could cover a broader topic or a larger number of topics than d_2, therefore having a larger number of unique terms that d_2. This is a consequence of not taking into account within document frequencies for terms.

Let us now consider the document rankings that RbLI could produce in response to a query q_1 from a collection with only the above two documents. In accordance with the imaging process, we will first perform imaging on document d_1, transfering all the probability of the not-d_1-terms to d_1-terms. Since d_1 has a large number of terms, then on average these terms will receive a small amount of probability from not-d_1-terms, and will have a posterior probability, after imaging, not much bigger than the their prior probability. The opposite will happen when we perform imaging on d_2. Since d_2 has a small number of terms, then on average these terms will receive a large amount of probability from not-d_2-terms, and will have a posterior probabil-

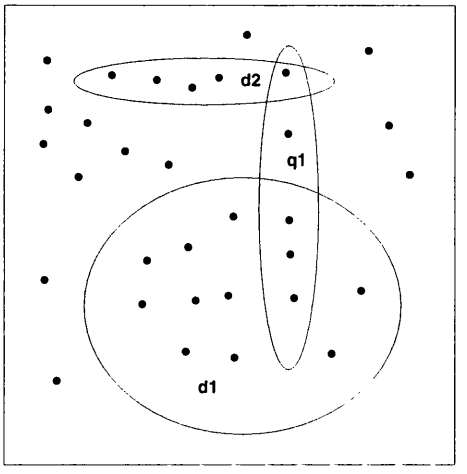

Figure 10.7: The document length effect.

ity, after imaging, much larger than the their prior probability. What will then happen is that document d_2 could be ranked in a higher position than d_1, even if the number of terms that q_1 and d_1 have in common is higher than the number of terms q_1 and d_2 have, as depicted in Figure 10.7.

The above effect is also present in the RbCP and in the RbGLI models, while it is absent from the RbJP model.

This undesired effect clearly suggests that some document length normalisation factor should be devised to avoid it to happen. However, the identification of the best normalisation strategy for both the RbLI model and the RbGLI models is a very complex issue that is outside the scope of this thesis. The identification of a normalisation strategy for these models will be left to future work.

10.6 Experiments with the RbGLI model

This section reports on experiments concerning the implementation and evaluation of the RbGLI model. This section is not as detailed as the previous section, since there was no need to repeat some of the experiments already carried out with the RbLI model if there was not reasons to believe they would have had different results with the RbGLI model. RbGLI is a generalisation of RbLI, but the kinematics of probabilities it causes is very similar and some of the results already achieved for the RbLI model can be easily

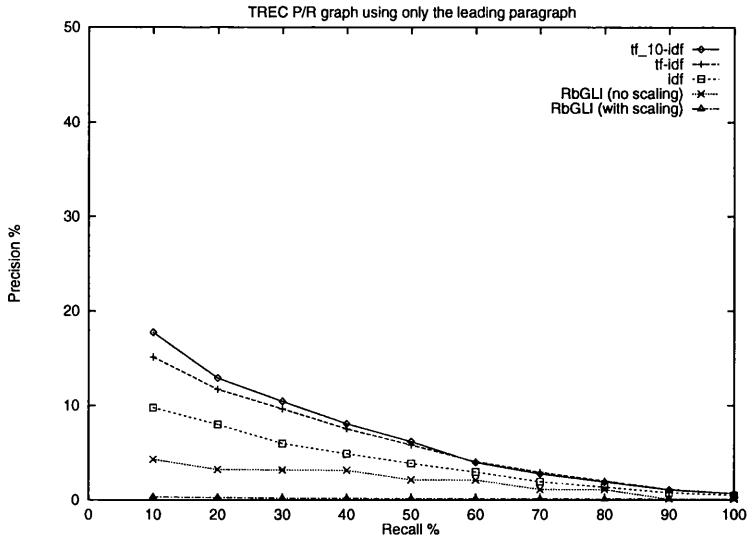

Figure 10.8: Precision and recall graphs for the WSJ-lead collection using the RbGLI model with or without probability scaling.

extended to RbGLI. Therefore, in studying the effectiveness of the RbGLI model a smaller number of experiments than in studying the RbLI model was carried out. This section only reports the most significant results.

10.6.1 Using only leading paragraphs

The experimentation of RbGLI was performed in the same way as the one for RbLI. Experiments were first performed on the WSJ-lead collection and on the larger WSJ-full collection. This section reports the results obtained for RbGLI using the WSJ-lead collection.

Redistributing the untransfered probabilities

Figure 10.8 reports the results of an experimental investigation into the effectiveness of performing a proportional transfer (scaling) of the probabilities remained untransfered after imaging, as suggested in Chapter 6.

The results show that RbGLI with scaling gives much worse performance than RbGLI without scaling. This result is in accordance with the one obtained for RbLI.

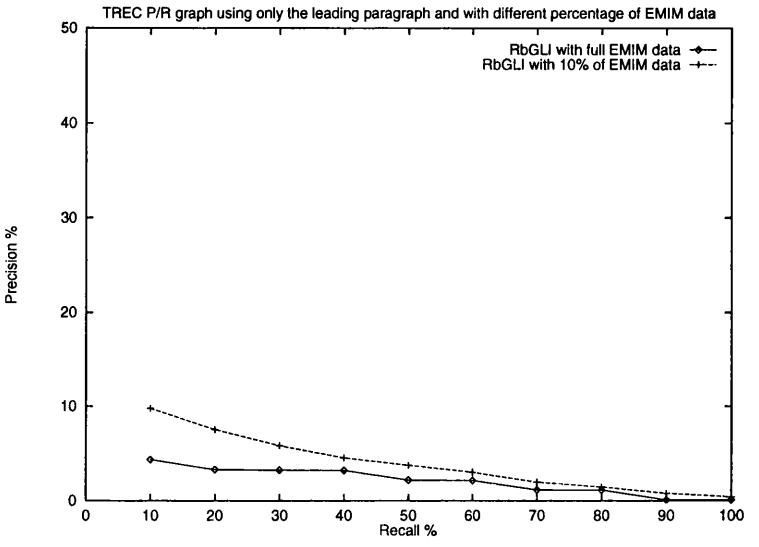

Figure 10.9: Precision and recall graphs for the WSJ-lead collection using different percentage of the full EMIM data.

Using different amounts of EMIM data

Figure 10.9 shows how the performance of the RbGLI model improves by using only a portion of the EMIM data. The 10% portion of the EMIM data proved to be the most effective, all other portions experimented (20%, 30%, and 40%) gave better results that using the full EMIM data, but worse that the 10% portion.

Again this result is not different from the one obtained using the RbLI model.

Using different stoplists

For the reasons already reported in the previous section regarding experimentation with RbLI, a set of experiments was performed to analyse the performance of RbGLI when terms with very high frequency of occurrence (and therefore with low *idf* weight) were removed from the term space.

Figure 10.10 shows that for RbGLI, as for RbLI, the use of a stoplist that includes the most frequent terms in the term space does improve performance. In particular, a stoplist made of standard stopterms and of the 1% most frequent terms in the term space proves to be the most effective one. This same stoplist proved to be the most effective also for the RbLI model.

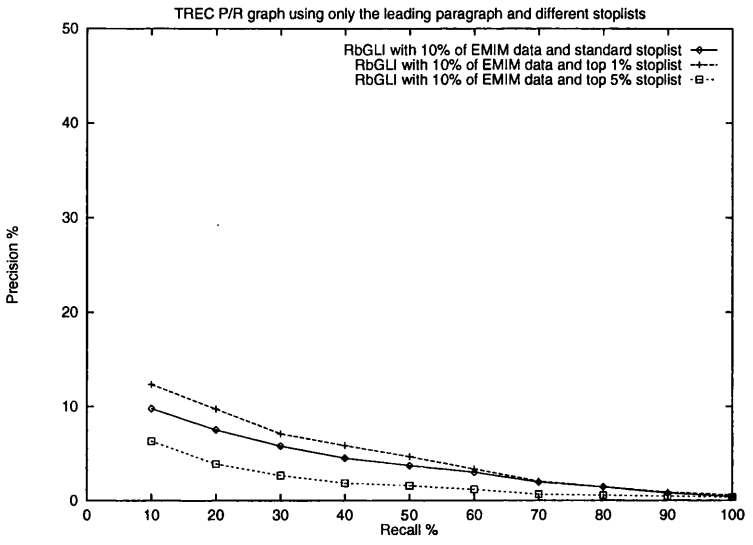

Figure 10.10: Precision and recall graphs for the WSJ-lead collection using different stoplists.

Choosing the best combination of parameters

In the light of the findings reported in the previous sections, the best combination of parameters for the RbGLI model proved to be the same as the one for the RbLI model. Figure 10.11 shows that using RbGLI with 10% of the EMIM data and a stoplist including the 1% most frequent terms gives performance that are better than those given by *idf* model. This result is very similar to the one obtained for the RbLI model, although here the difference in performance between *idf* and RbGLI seems larger.

Figure 10.11 shows also a comparison between the performance of RbLI and RbGLI using their best combination of parameters. RbGLI performs almost exactly as RbLI. This result is rather disappointing and dissimilar to that obtained for smaller test collections (see Chapter 4).

10.6.2 Using full documents

This section reports the results obtained experimenting RbGLI using the WSJ-full collection.

Figure 10.12 reports a performance comparison between RbGLI, RbLI, and some classical IR models. As it can be seen RbGLI, RbLI, and *idf* have

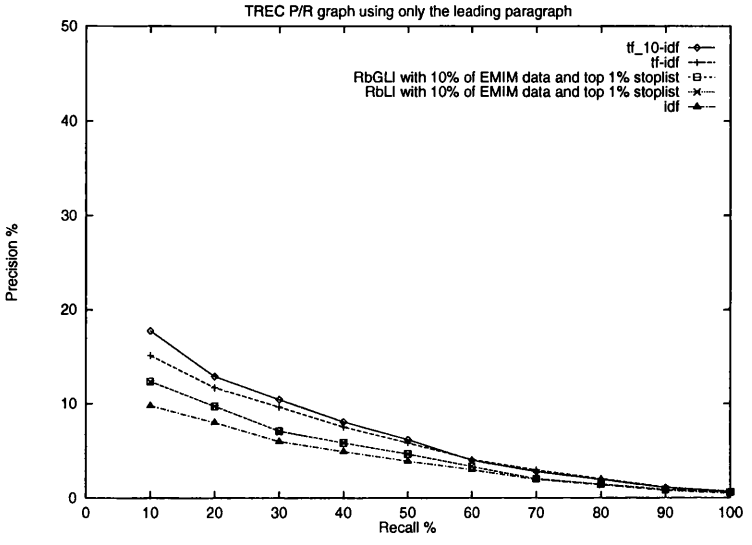

Figure 10.11: Performance of the RbGLI model using the WSJ-lead collection.

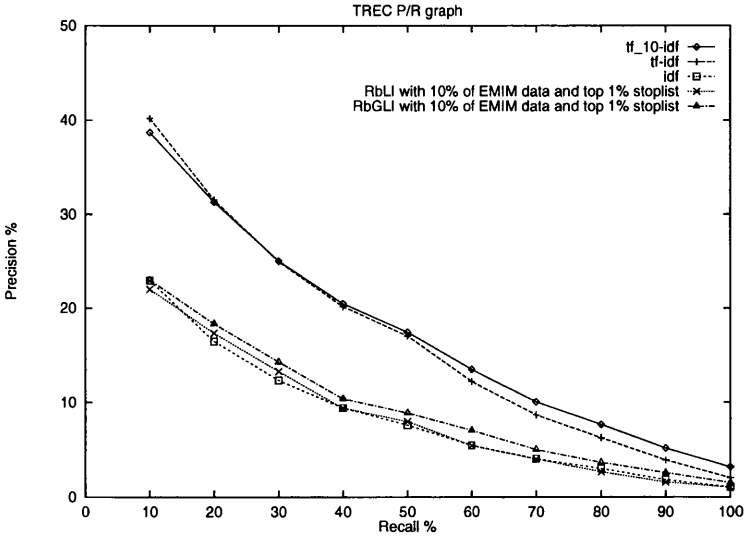

Figure 10.12: Precision and recall graph for the WSJ-full collection.

basically the same levels of performance. The $tf_{10} - idf$ and the $tf - idf$ also have basically the same levels of performance, but higher than the RbGLI, RbLI, or *idf* ones.

Data sets:	*WSJ-full*	*WSJ-lead*
num. of documents	74.520	74.520
size in MB	247	72
unique terms in documents	123.852	61.079
avg. doc. length	180	60
std. dev. of doc. length	142,64	26,38

Table 10.2: Average and standard deviation of the number of unique terms in WSJ-full and WSJ-lead.

The reason why RbGLI and RbLI perform better that *idf* using WSJ-lead collection, but have basically the same levels of performance than *idf* for the WSJ-full collection, is to be found in the document length effect discussed previously. Documents in the WSJ-lead collection have a more homogeneous length than documents in the WSJ-full collection. This is obvious, since the leading paragraphs of newspaper articles does not vary as much in length as the articles themselves. Therefore the document length effect plays a more important role in the WSJ-full collection than in WSJ-lead collection, decreasing the effectiveness of RbLI and RbGLI. This hypothesis is confirmed by the data reported in Table 10.2, that show the average and standard deviation of the number of unique terms in WSJ-lead and WSJ-full.

10.7 Comparison with the results obtained using smaller test collections

It is interesting to compare the results reported here and obtained using the WSJ document collection with those reported in Chapters 3 and 4 obtained using smaller test collections. In fact, these results do not seem to agree completely.

Figures 4.7, 4.8, and 4.9 report a comparison of the performance of RbJP, RbCP, RbLI and RbGLI using three different small test collections: the CACM, the Cranfield, and the NPL. In those figures, RbLI and RbGLI perform significantly better than RbJP. We could therefore conclude that the use of similarity information between terms in directing the kinematics

of probabilities in the term space helps improving performance. Considering now that RbJP is just another name for the classical *idf* model, if we compare the results of Figures 4.7, 4.8, and 4.9 with those reported in Figures 10.5, 10.6, 10.11, and 10.12, we can see that the above conclusion seems to lose strength. The difference in performance between RbJP, RbLI and RbGLI using a large collection of document does not seem significant enough to support that conclusion anymore.

It is interesting to note that the difference in performance between RbJP, RbLI and RbGLI seems to become smaller with the increasing of the size of the collection and therefore of the term space. It is difficult to find an explanation for this effect. Other different characteristics of the collections used in the experimentation, a part from size, could effect this result. Surely the document length effect has a big part in the different results obtained, as the use of EMIM as a measure of semantic similarity between terms. An in depth study on how different characteristics of the term space and different metrics on the space influence the kinematics of probability will be the task of future research.

10.8 Conclusions

In this chapter the effectiveness of the RbLI and the RbGLI models has been tested using a large collection of documents. Two different collections having different characteristics of the term space were constructed from the original WSJ collection. RbLI and RbGLI has been tested using both these collections against classical models of IR, like the *idf* model (corresponding to the RbJP model of Chapter 4) and the $tf - idf$ model.

Although, taken on their own, these results seem disappointing, some explanations for the low levels of performance have been found. An understanding of the factors causing these low performance does certainly help in suggesting ways to improve them. Therefore, the major contribution of the results reported in this chapter is in providing indications for future work, as it will be discussed in the next chapter.

Part VI

Conclusions

Chapter 11

Conclusions and Future Work

This chapter summarises the theoretical and experimental contributions of the work reported in this thesis. It also addresses the limitations of these contributions and of the approach followed. The chapter ends with directions for future work.

11.1 Conclusions

This thesis studied the kinematics of probabilities in probabilistic IR. The aim was to get a better insight of the behaviour of the probabilistic models of IR currently in use and to propose new and more effective models by exploiting different kinematics of probabilities. The study was performed both from a theoretical and an experimental point of view. In the following the conclusions of this work are summarised, distinguishing between theoretical and experimental conclusions.

11.1.1 Theoretical conclusions

The theoretical conclusions of the study of the probability kinematics in IR reported in this thesis, and in particular in Chapters 3 and 4, show that a probability transfer between terms in the term space that takes into account the semantic similarity between the probability-donor term and the probability-recipient term is more effective in the context of IR than a probability transfer that does not take that into account. Most current proba-

bilistic retrieval models are based on a probability kinematics that does not take into account similarity between terms or between documents, unless "ad hoc" weighting schemas, mostly based on clustering, are used.

This result was achieved by:

1. considering a term space for which similarity information between terms could be obtained;
2. designing two new theoretical IR models whose probability kinematics mimic those of classical IR models, like the vector space model and the probabilistic model;
3. designing two new theoretical models of IR whose probability kinematics take into account the similarity between terms in the term space;
4. comparing the different behaviour and effectiveness of the four models.

Although this result seems to be supported by the experiments performed on small test collections reported in Chapters 3 and 4, it should be stressed that this is a purely theoretical result. In fact, the two new models (RbLI and RbGLI) were compared with two fictitious models (RbJP and RbCP) that, although having a probability kinematics similar to the vector space and the probabilistic models, have not been optimised for effectiveness with ad hoc parameters. The comparison between the four models was on a very simplified testing ground.

Nevertheless, the above result suggests the usefulness of a further investigation into more complex and optimised models for probabilistic retrieval, where probability kinematics follows non-classical approaches. The RbLI and RbGLI models proposed in this thesis are just two of such approaches, but others can be developed using results achieved in other fields, such as for example Conditional Logic, Modal Logic, and Belief Revision theory.

11.1.2 Experimental conclusions

The theoretical results summarised in the previous section suggest that an improvements in retrieval effectiveness can be obtained by designing probabilistic IR systems that are based upon a probability kinematics that exploit semantic similarity between terms in the term space. Unfortunately, while experiments using small test collections seem to provide evidence supporting

this conclusion, experiments performed using large test collections do not seem to provide as much supporting evidence (although they do not seem to provide contrasting evidence neither). The peculiar characteristics of the term space of different collections play an important role in shaping the effects that different probability kinematics have in the effectiveness of the retrieval process. Characteristics such as the size of the term space, term frequency, document length, and term co-occurrence have effects that are difficult to factorise in a experimental study of retrieval effectiveness.

A much larger experimental investigation than the one reported in this thesis is necessary in order to find out the influence that each characteristic of the collections has in the kinematics of term weights.

11.2 Limitations and future work

There are a number of limitations to the work reported in this thesis. The following are the most important.

Factors influencing the kinematics of probabilities in probabilistic IR

The most important limitation of the work reported in this thesis is that the study only analyses the influence on retrieval effectiveness of two factors:

1. the term distribution in the term space;
2. the semantic similarity between terms.

Term weights were determined from the term distribution in the term space and their kinematics at retrieval time was studied. New models that could make use of semantic similarity between terms in directing the kinematics of term weights at retrieval time were proposed. The effectiveness of these new models was compared with that of classical IR models that only take into consideration the term distribution.

Although there is some experimental evidence supporting the fact that these new models improve effectiveness, experimental result confirmed that two other factors need to be taken into consideration for the design of effective probabilistic IR systems:

3. the document length;

4. the term distribution inside the document.

Future research will have to look into ways of including these two factors into the new models proposed. This will require both theoretical and experimental work.

Probability distribution on the term space

In the study reported in this thesis the *idf* weight was used as a measure of the importance of a term in the context of the term space. Although this seems a reasonable choice in the context of IR, it should be noticed that the *idf* weight has a semantic interpretation that is in contrast with the usual semantic interpretation given to probability. The probability of an event can have two semantic interpretations, as Carnap [Car50] pointed out: a frequentist interpretation and a belief interpretation. Without entering into philosophical considerations, the *idf* weight is in contrast with both these semantics, being proportional to the rareness of the occurrence of a term in the term space, and not to its frequency of occurrence. Basically, a term with high *idf* weight has a low probability of occurrence; therefore, although the *idf* weight can be considered as a good measure of the importance of a term in the term space, we should consider that the theory of imaging [G̈82, Lew86] was devised with the concept of probability in mind, not its opposite.

It will be necessary in the future to investigate theoretically if there is any contradiction in the use this thesis made of the imaging theory.

Semantic similarity between terms

In this thesis, the semantic similarity between terms was estimated using EMIM. The identification of the most appropriate measure of semantic similarity is the major requirement for the definition of a metrics on the term space. Such a metric provides a way of determining the accessibility relation between terms that is at the core of the imaging process.

However, EMIM was here estimated using the technique proposed by Van Rijsbergen in [vR79]. This technique makes use of occurrence information for terms in the term space, and estimates the similarity between terms making

a heavy use of the frequency of co-occurrence of terms in the space. Because of the way EMIM is evaluated, there are two problems with its use in the context of the imaging theory that need further investigation:

- the data used to evaluate the semantic similarity between terms is collection dependent, that is to say, the semantic similarity between terms depends on how the terms are used in the documents of the collection and can be different from one collection to another;
- EMIM is estimated using term occurrence information and so is the *idf* weight; we are therefore using the same information to estimate the relative importance of a term in the term space and its semantic similarity with other terms.

Future research will have to look for other measures of similarity between terms upon which to build a metric on the term space. These measures should be collection independent and should be more oriented to the semantics of terms than to their distributions. The use of a thesaurus, for example, seems to be the next obvious direction to follow.

Document length normalisation

In the work reported in this thesis the document length (measured as number of terms occurring in the document) was not used for normalisation purposes. Document length normalisation is a very important component of any retrieval model; without document length normalisation factors most ranked retrieval models would rank the longest documents first (see Chapter 2).

The negative effect that document length plays in the models proposed in this thesis has already been address in Chapter 10. It will be necessary in the future to modify the RbLI and the RbGLI models to take into account document length. This work is both theoretical and experimental, since the theory of imaging does not provide any help for this issue and since any proposed modification to the above models will require supporting experimental evidence.

Document dependent information

Most classical probabilistic models of IR take into consideration not only the distribution of terms in the term space, but also the distribution of term in the context of single documents (see Chapter 2). The tf weight, used in the $tf - idf$ model, is a classical example of use of document dependent information. The importance of a term, represented by the weight associated to the term $(tf - idf)$, is therefore composed of two parts: a collection wide one (idf), measuring the importance of a terms in the term space, and a document dependent one (tf), measuring the importance of a term in the context of the document under consideration.

In the work reported in this thesis document dependent information was not taken into consideration. Future research will have to investigate the best possible way of using document dependent information together with the other sources of information that the models based on the imaging theory already use. Again, this work will be both theoretical and experimental in nature.

Further experimentation

Future experimental work will have to investigate the above issues using different test collections, analysing the influence the characteristics of the collections have on the effectiveness of the theoretical proposals.

The methodology and the results reported in this thesis provide a very good starting point for the future work suggested above. A Ph.D. is never really finished!

Part VII

Bibliography

[AK92] G. Amati and S. Kerpedjiev. An Information Retrieval logical model: implementation and experiments. Technical Report Rel 5B04892, Fondazione Ugo Bordoni, Roma, Italy, March 1992.

[AvR95] G. Amati and C.J. van Rijsbergen. Probability, information and information retrieval. In *Proceedings of the First International Workshop on Information Retrieval, Uncertanty and Logic*, Glasgow, Scotland, UK, September 1995.

[BC76] A. Bookstein and W.S. Cooper. A general mathematical model for information retrieval systems. *The Library Quarterly*, 46(2), 1976.

[BC89] C. Berrut and Y. Chiaramella. Indexing medical reports in a multimedia environment: the RIME experimental approach. In *Proceedings of ACM SIGIR*, pages 187–197, Cambridge, MA, USA, June 1989.

[BDPd+92] P.F. Brown, V.J. Della Pietra, P.V. deSouza, J.C. Lai, and R.L. Mercer. Class-based n-gram models of natural language. *Computational Linguistics*, 18(4):467–479, 1992.

[BE90] J. Barwise and J. Etchemendy. *The language of First-Order Logic.* CSLI, Menlo Park, CA, USA, 1990.

[BFK+88] P. Biebricher, N. Fuhr, G. Knorz, G. Lustig, and M. Schwantner. The automatic indexing system AIX/PHYS - from research to application. In *Proceedings of ACM SIGIR*, pages 333–342, Grenoble, France, 1988.

[BL96] T. Berners-Lee. WWW: past, present, and future. *IEEE Computer*, October:69–77, 1996.

[Bla96] D.C. Blair. STAIRS Redux: thoughts on the STAIRS evaluation, ten years after. *Journal of the American Society for Information Science*, 47(1):4–22, 1996.

[Blo97] M. Bloemer. Evaluierung von retrieval-strategien mittels probabilistischem datalog. Diploma Thesis, University of Dortmund, Germany, Informatik VI, August 1997.

[BP93] G. Borgogna and G. Pasi. A fuzzy linguistic approach generalizing boolean information retrieval: a model and its evaluation. *Journal of the American Society for Information Science*, 2(70-82):44, 1993.

[Bru93] P.D. Bruza. *Stratified Information Disclosure: a synthesis between Hypermedia and Information Retrieval.* Phd thesis, Katholieke Universiteit Nijmegen, The Netherlands, 1993.

[BS74] A. Bookstein and D. Swanson. Probabilistic models for automatic indexing. *Journal of the American Society for Information Science*, 25(5):312–318, 1974.

[Bv92] P.D. Bruza and T.P. van der Weide. Stratified hypermedia structures for information disclosure. *The Computer Journal*, 35(3):208–220, 1992.

[Car50] R. Carnap. *Logical Foundations of probability.* Routledge and Kegan Paul Ltd, London, UK, 1950.

[CC92] Y. Chiaramella and J.P. Chevallet. About retrieval models and logic. *The Computer Journal*, 35(3):233–242, 1992.

[CGD92] W.S. Cooper, F.C. Gey, and D.P. Dabney. Probabilistic retrieval based on staged logistic regression. In *Proceedings of ACM SIGIR*, pages 198–210, Copenhagen, Denmark, June 1992.

[CGG92] J. Cowie, J. Guthrie, and L. Guthrie. Lexical disambiguation using simulated annealing. In *Proceedings of the COLING Conference*, pages 359–365, August 1992.

[CH79] W.B. Croft and D.J. Harper. Using probabilistic models of document retrieval without relevance information. *Journal of Documentation*, 35:285–295, 1979.

[CH89] K.W. Church and P. Hanks. Word association norms, mutual information and lexicography. In *Proccedings of ACL 27*, pages 76–83, Vancouver, Canada, 1989.

[Chu92] K.W. Church. Using bilingual materials to develop word sense disambiguation methods. In *Proceedings of ACM SIGIR*, page 350, Copenhagen, Denmark, June 1992.

[CLC88] W.B. Croft, T.J. Lucia, and P.R. Cohen. Retrieving documents by plausible inference: a preliminary study. In *Proceedings of ACM SIGIR*, Grenoble, France, June 1988.

[CLCW89] W.B. Croft, T.J. Lucia, J. Crigean, and P. Willet. Retrieving documents by plausible inference: an experimental study. *Information Processing and Management*, 25(6):599–614, 1989.

[CMK66] C. Cleverdon, J. Mills, and M. Keen. *ASLIB Cranfield Research Project: factors determining the performance of indexing systems.* ASLIB, 1966.

[Coo71] W.S. Cooper. A definition of relevance for information retrieval. *Information Storage and Retrieval*, 7:19–37, 1971.

[Coo94] W.S. Cooper. The formalism of probability theory in IR: a foundation or an encumbrance. In *Proceedings of ACM SIGIR*, pages 242–247, Dublin, Ireland, June 1994.

[Coo95] W.S. Cooper. Some inconsistencies and misnomers in probabilistic information retrieval. *ACM Transactions on Information Systems*, 13(1):100–111, 1995.

[Cox70] D.R. Cox. *Analysis of Binary Data.* Methuen, London, UK, 1970.

[CR87] W.B. Croft and R.H.Thompson. I^3R: a new approach to the design of Document Retrieval Systems. *Journal of the American Society for Information Science*, 38(6):389–404, 1987.

[CR95] F. Crestani and T. Roelleke. Issues on the implementation of imaging on top of probabilistic datalog. In *Proceedings of the First Workshop in IR, Uncertainty and Logic.* Glasgow, Scotland, UK, September 1995.

[Cro87] W.B. Croft. Approaches to Intelligent Information Retrieval. *Information Processing and Management*, 23(4):249:254, 1987.

[Cro92] W.B. Croft. The University of Massachusetts TIPSTER project. In *Proceeding of the TREC Conference.* Gaithersburg, MD, USA, November 1992.

[Cro94] C.B. Cross. Eliminative bayesianism and probability revision. Unpublished paper, August 1994.

[CRSvR95] F. Crestani, I. Ruthven, M. Sanderson, and C.J. van Rijsbergen. The troubles with using a logical model of IR on a large collection of documents. Experimenting retrieval by logical imaging on TREC. In *Proceeding of the TREC Conference*, Washington D.C., USA, November 1995.

[CST92] W.B. Croft, L.A. Smith, and H.R. Turtle. A loosely-coupled integration of a text retrieval system and an object-oriented database system. In *Proceedings of ACM SIGIR*, pages 223–232, Copenhagen, Denmark, June 1992.

[CSTL97] F. Crestani, M. Sanderson, M. Theophylactou, and M. Lalmas. Short queries, natural language, and spoken document retrieval: experiments at glasgow university. In *Proceeding of the TREC Conference*, Washington D.C., USA, November 1997. (In press).

[CSv96] F. Crestani, M. Sanderson, and C.J. van Rijsbergen. Sense resolution properties of logical imaging. *The New Review of Document and Text Management*, 1:277–298, 1996.

[CSvR96] F. Crestani, F Sebastiani, and C.J. van Rijsbergen. Imaging and information retrieval: variation on a theme. In *Proceedings of the Second International Workshop on Information Retrieval, Uncertainty and Logic*, pages 27–31, Glasgow, UK, July 1996.

[Cv95] F. Crestani and C.J. van Rijsbergen. Probability kinematics in information retrieval. In *Proceedings of ACM SIGIR*, pages 291–299, Seattle, WA, USA, July 1995.

[CvR95] F. Crestani and C.J. van Rijsbergen. Information Retrieval by Logical Imaging. *Journal of Documentation*, 51(1):1–15, 1995.

[CvR96] I. Campbell and C.J. van Rijsbergen. The ostensive model of developing information needs. In *Proceedings of CoLIS 2*, pages 251–268, Copenhagen, Denmark, October 1996.

[Dem68] A. P. Dempster. A generalization of the Bayesian inference. *Journal of Royal Statistical Society*, 30:205–447, 1968.

[Dem93] G.C. Demetriou. Lexical disambiguation using constraint handling in prolog (chip). In *Proceedings of the European Chapter of the ACL*, pages 431–436, 1993.

[DIS91] I. Dagan, A. Itai, and U. Schwall. Two languages are more informative than one. In *Proceedings of the ACL*, pages 130–137, 1991.

[dSM93] W. Teixera de Silva and R. L. Milidiu. Belief function model for information retrieval. *Journal of the American Society of Information Science*, 4(1):10–18, 1993.

[Dun91] M.D. Dunlop. *Multimedia Information Retrieval.* PhD Thesis, Department of Computing Science, University of Glasgow, Glasgow, UK, October 1991.

[Eft96] E. Efthimiadis. Query expansion. *Annual Review of Information Science and Technology*, 31:121–187, 1996.

[FB91] N. Fuhr and C. Buckley. A probabilistic learning approach for document indexing. *ACM Transactions on Information Systems*, 9(3):223–248, 1991.

[FB93] N. Fuhr and C. Buckley. Optimizing document indexing and search term weighting based on probabilistic models. In D. Harman, editor, *The First Text Retrieval Conference (TREC-1)*, Special Publication 500-207, pages 89–100, Gaithersburg, MD, USA, 1993. National Institute of Standareds and Technology.

[FBY92] W.R. Frakes and R. Baeza-Yates, editors. *Information Retrieval: data structures and algorithms.* Prentice Hall, Englewood Cliffs, New Jersey, USA, 1992.

[FCAT90] R.M. Fung, S.L. Crawford, L.A. Appelbaum, and R.M. Tong. An architecture for probabilistic concept based information retrieval. In *Proceedings of ACM SIGIR*, pages 455–467, Bruxelles, Belgium, September 1990.

[Fei85] S.J. Feinglos. *MEDLINE: a basic guide to searching.* Medical Library Association, Chicago, IL, USA, 1985.

[FK84] N. Fuhr and G. Knowrz. Retrieval test evaluation of a rule based automatic indexing (AIR/PHYS). In C.J. van Rijsbergen, editor, *Research and development in Information Retrieval*, pages 391–408. Cambridge University Press, Cambridge, UK, 1984.

[FP91] N. Fuhr and U. Pfeifer. Combining model-oriented and description-oriented approaches for probabilistic indexing. In

Proceedings of ACM SIGIR, pages 46–56, Chicago, USA, October 1991.

[FR95] N. Fuhr and T. Rölleke. A probabilistic relational algebra for the integration of information retrieval and database systems. Submitted for publication, 1995.

[FR97] N. Fuhr and T. Roelleke. Hyspirit: a probabilistic inference engine for hypermedia retrieval in large databases. Submitted for publication, November 1997.

[Fuh89] N. Fuhr. Models for retrieval with probabilistic indexing. *Information Processing and Management*, 25(1):55–72, 1989.

[Fuh90] N. Fuhr. A probabilistic framework for vague queries and imprecise information in databases. In *Proceedings of the International Conference on Very Large Databases*, pages 696–707, Los Altos, CA, USA, 1990. Morgan Kaufman.

[Fuh92a] N. Fuhr. Integration of probabilistic fact and text retrieval. In *Proceedings of ACM SIGIR*, pages 211–222, Copehagen, Denmark, June 1992.

[Fuh92b] N. Fuhr. Probabilistic models in Information Retrieval. *The Computer Journal*, 35(3):243–254, 1992.

[Fuh93] N. Fuhr. A probabilistic relational model for the integration of IR and Databases. In *Proceedings of ACM SIGIR*, pages 309–317, Pittsburgh, PA, USA, June 1993.

[Fuh95] N. Fuhr. Probabilistic datalog - a logic for powerful retrieval methods. In *Proceedings of ACM SIGIR*, pages 282–290, Seattle, WA, USA, July 1995.

[Gä82] P. Gärdenfors. Imaging and conditionalization. *Journal of Philosophy*, 79:747–760, 1982.

[Gä88] P. Gärdenfors. *Knowledge in flux: modelling the dynamics of epistemic states.* The MIT Press, Cambridge, Massachusetts, USA, 1988.

[GCY92] W. Gale, K.W. Church, and D. Yarowsky. Estimating upper and lower bounds on the performance of word-sense disambiguation programs. In *Proceedings of the ACL*, pages 249–256, 1992.

[Goo50] I. J. Good. *Probability and the Weighing of Evidence.* Charles Griffin Symand Company Limited, 1950.

[Hal90] J.Y. Halpern. An analysis of first-order logics for probability. *Artificial Intelligence*, 46:311–350, 1990.

[Har75] S.P. Harter. A probabilistic approach to automatic keyword indexing: part 1. *Journal of the American Society for Information Science*, 26(4):197–206, 1975.

[Har92a] D. Harman. Ranking algorithms. In W.B. Frakes and R. Baeza-Yates, editors, *Information Retrieval: data structures and algorithms*, chapter 14. Prentice Hall, Englewood Cliffs, New Jersey, USA, 1992.

[Har92b] D. Harman. Relevance feedback and other query modification techniques. In W.B. Frakes and R. Baeza-Yates, editors, *Information Retrieval: data structures and algorithms*, chapter 11. Prentice Hall, Englewood Cliffs, New Jersey, USA, 1992.

[Har92c] D. Harman. Relevance feedback revisited. In *Proceedings of ACM SIGIR*, pages 1–10, Copenhagen, Denmark, June 1992.

[Har93] D. Harman. Overview of the first TREC conference. In *Proceedings of ACM SIGIR*, pages 36–47, Pittsburgh, PA, USA, June 1993.

[Har94] D. Harman. Overview of the third text retrieval conference (TREC-3). In *Proceeding of the TREC Conference*, Gaithersburg, MD, USA, November 1994.

[Har95a] D. Harman. Overview of the fourth text retrieval conference (TREC-4). In *Proceeding of the TREC Conference*, Gaithersburg, MD, USA, November 1995.

[Har95b] D. Harman. Overview of the second text retrieval conference (TREC-2). *Information Processing and Management*, 31(3):271–289, 1995.

[Har96] D. Harman. Overview of the fifth text retrieval conference (TREC-5). In *Proceeding of the TREC Conference*, Gaithersburg, MD, USA, November 1996.

[HC68] G.E. Hughes and M.K. Cresswell. *An Introduction to Modal Logic.* Muthuen and Co. Ltd, London, UK, 1968.

[HLvR96] T.W.C. Huibers, M. Lalmas, and C.J. van Rijsbergen. Information retrieval and situation theory. *SIGIR Forum*, 30(1):11–25, 1996.

[Hui96] T.W.C Huibers. *An Axiomatic Theory for Information Retrieval.* PhD thesis, Utrecht University, The Netherlands, 1996.

[Hul88] J. Hullman. *Principles of Database and Knowledge-Base Systems*, volume I. Computer Science Press, Rockville, MD, USA, 1988.

[HW92] D.J. Harper and A.D.M. Walker. ECLAIR: an extensible class library for Information Retrieval. *The Computer Journal*, 35(3):256–267, 1992.

[Jef65] R.C. Jeffrey. *The logic of decision.* McGraw-Hill, New York, USA, 1965.

[KC87] R. Kjeldsen and P.R. Cohen. The evolution and performance of the GRANT system. *IEEE Expert*, summer:73–79, 1987.

[KD88] B.W. Kernighan and Ritchie D.M. *The C programming language.* Prentice Hall, Englewood Cliffs, NJ, USA, second edition, 1988.

[Kri71] S.A. Kripke. Semantical considerations on modal logic. In L. Linsky, editor, *Reference and modality*, chapter 5, pages 63–73. Oxford University Press, Oxford, UK, 1971.

[KS75] E. Kelly and P. Stone. *Computer recognition of english word senses.* North-Holland Publishing Co, Amsterdam, 1975.

[Kwo90] K.L. Kwok. Experiments with a component theory of probabilistic Information Retrieval based on single terms as document components. *ACM Transactions on Information Systems*, 8(4):363–386, 1990.

[Lal92] M. Lalmas. A logic model of information retrieval based on situation theory. In *Proccedings of the 14th BCS Information Retrieval Colloquium*, Lancaster, UK, December 1992.

[Lal96] M. Lalmas. *Theories of Information and Uncertainty for the modelling of Information Retrieval: an application of Situation Theory and Dempster-Shafer 's Theory of Evidence.* PhD thesis, Department of Computing Science, University of Glasgow, Glasgow, Scotland, UK, 1996.

[Lal97] M. Lalmas. Logical models in information retrieval: introduction and overview. *Information Processing and Management*, 1997. In print.

[Les86] M. Lesk. Automatic sense disambiguation: how to tell a pine cone from an ice cream cone. In *Proceedings of the SIGDOC Conference*, pages 24–26, June 1986.

[Lew81] D. Lewis. Probability of conditionals and conditionals probabilities. In W.L. Harper, R. Stalnaker, and G. Pearce, editors, *Ifs*, The University of Western Ontario Series in Philosophy of Science, pages 129–147. D.Reidel Publishing Company, Dordrecht, Holland, 1981.

[Lew86] D. Lewis. *Conterfactuals.* Basil Blackwell, Oxford, UK, 2nd edition, 1986.

[Luh57] H.P. Luhn. A statistical approach to mechanized encoding and searching of library Information. *IBM Journal of Research and Development*, 1:309:317, 1957.

[LvR93] M. Lalmas and C.J. van Rijsbergen. A model of an Information Retrieval system based on Situation Theory and Dempster-Shafer theory of evidence. In *Proceedings of the 1st Workshop on Incompleteness and Uncertainty in Information Systems*, pages 62–67, Montreal, Canada, 1993.

[Mar92] E.L. Margulis. N-poisson document modelling. In *Proceedings of ACM SIGIR*, pages 177–189, Copenhagen, Denmark, June 1992.

[Mar93] E.L. Margulis. Modelling documents with multiple Poisson distributions. *Information Processing and Management*, 29(2):215–227, 1993.

[Men95] F. Menczer. Internet: vita da ragni. *Sistemi Intelligenti*, 7(3):421–442, 1995.

[Mil71] W.L. Miller. A probabilistic serach strategy for MEDLARS. *Journal of Documentation*, 27:254–266, 1971.

[Miz96] S. Mizzaro. Relevance: the whole (hi)story. Technical Report UDMI/12/96/RR, Dipartimento di Matematica e Informatica, Universita' di Udine, Italy, December 1996.

[MK60] M.E. Maron and J.L. Kuhns. On relevance, probabilistic indexing and retrieval. *Journal of the ACM*, 7:216–244, 1960.

[Nea90] R.E. Neapolitan. *Probabilistic reasoning in expert systems.* John Wiley and Son Inc., New York, USA, 1990.

[Nie88] J.Y. Nie. An outline of a general model for information retrieval. In *Proceedings of ACM SIGIR*, pages 495–506, Grenoble, France, June 1988.

[Nie89] J.Y. Nie. An Information Retrieval model based on Modal Logic. *Information Processing and Management*, 25(5):477–491, 1989.

[Nie92] J.Y. Nie. Towards a probabilistic modal logic for semantic based information retrieval. In *Proceedings of ACM SIGIR*, pages 140–151, Copenhagen, Denmark, June 1992.

[Nil86] N.J. Nilsson. Probabilistic logic. *Artificial Intelligence*, 28:71–87, 1986.

[NLB95] J.Y. Nie, F. Lepage, and M. Brisebois. Information retrieval as counterfactuals. *The Computer Journal*, 38(8):643–657, 1995.

[Nut80] D. Nute. *Topics in Conditional logic.* D. Reidel Publishers. Dodrecht, 1980.

[OB91] R.N. Oddy and B. Balakrishnan. Pthomas: and adaptive Information Retrieval system on the Connection Machine. *Information Processing and Management*, 27(4):317–335, 1991.

[Pea88] J. Pearl. *Probabilistic reasoning in intelligent systems: networks of plausible inference.* Morgan Kaufmann, San Mateo, California, 1988.

[Pea90] J. Pearl. Jeffrey's rule, passage of experience and Neo-Bayesianism. In H.E. Kyburg, R.P. Luoi, and G.N. Carlson, editors, *Knowledge representation and defeasible reasoning*, pages 245–265. Kluwer Academic Publisher, Dodrecht, The Netherlands, 1990.

[Por80] M.F. Porter. An algorithm for suffix stripping. *Program*, 14(3):130–137, 1980.

[QF93] Y. Qiu and H.P. Frei. Concept based query expansion. In *Proceedings of ACM SIGIR*, pages 160–171, Pittsburgh, PA, USA, June 1993.

[Rö95] T. Rölleke. Imaging on top of probabilistic datalog. In *Proceedings of ACM SIGIR*, Seattle, WA, USA, July 1995. Poster session.

[RMC82] S. E. Robertson, M. E. Maron, and W. S. Cooper. Probability of relevance: a unification of two competing models for document retrieval. *Information Technology: Research and Development*, 1:1–21, 1982.

[Rob76] S.E. Robertson. *A theoretical model of the retrieval characteristics of information retrieval systems.* PhD Thesis, University of London, UK, 1976.

[Rob77] S.E. Robertson. The probability ranking principle in IR. *Journal of Documentation*, 33(4):294–304, December 1977.

[RS76] S.E. Robertson and K. Sparck Jones. Relevance weighting of search terms. *Journal of the American Society for Information Science*, 27:129–146, May 1976.

[RW94] S.E. Robertson and S. Walker. Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval. In *Proceedings of ACM SIGIR*, pages 232–241, Dublin, Ireland, June 1994.

[Sal68] G. Salton. *Automatic information organization and retrieval.* Mc Graw Hill, New York, 1968.

[Sal71] G. Salton. *The SMART Retrieval System. Experiments in automatic document processing.* Prentice-Hall, New Jersey, 1971.

[San89] D.H. Sanford. *If P, then Q: conditionals and the foundations of reasoning.* Routledge, London, UK, 1989.

[San96a] M. Sanderson. System for information retrieval experiments (SIRE). Unpublished paper, November 1996.

[San96b] M. Sanderson. *Word Sense Disambiguation and Information Retrieval.* PhD Thesis, Department of Computing Science, University of Glasgow, Glasgow, Scotland, UK, 1996.

[Sav92] J. Savoy. Bayesian inference networks and spreading activation in hypertext systems. *Information Processing and Management*, 28(3):389–406, 1992.

[Seb94] F. Sebastiani. A probabilistic terminological logic for modelling information retrieval. In *Proceedings of ACM SIGIR*, pages 122–131, Dublin, Ireland, 1994.

[Seb96] F. Sebastiani. Information retrieval, imaging and probabilistic logic. In *Deliverable D3: A Theory of Uncertainty for Information Retrieval*, number 1/96 in FERMI, technical report 7, pages 59–65. ESPRIT Basic Research Action, Project Number 8134 - FERMI, March 1996.

[Ser70] T. Seracevic. The concept of "relevance" in information science: a historical review. In T. Seracevic, editor, *Introduction to Information Science*, chapter 14. R.R. Bower Company, New York, USA, 1970.

[SH93] S. S. Schoken and R. A. Hummel. On the use of dempster shafer model in information indexing and retrieval applications. *International Journal of Man-Machine Studies*, 39:1–37, 1993.

[Sha76] G. Shafer. *A Mathematical Theory of Evidence.* Princeton University Press, 1976.

[SJ81] K. Sparck Jones. *Information Retrieval Experiments.* Butterworth, London, 1981.

[SJ95] K. Sparck Jones. Reflections on TREC. *Information Processing and Management*, 31(3):291–314, 1995.

[SR82] S. Small and C. Rieger. *Strategies for Natural Language Processing.* LEA, 1982.

[SR96] M. Sanderson and I. Ruthven. Report of the glasgow ir group submission. In *Proceedings of the Fifth Text Retrieval Conference (TREC-5)*, Washington D.C., USA, November 1996.

[Sri92] P. Srinivadsan. Thesaurus construction. In W.B. Frakes and R. Baeza-Yates, editors, *Information Retrieval: data structures and algorithms.*, chapter 9. Prentice Hall, Englewood Cliffs, New Jersey, USA, 1992.

[SS88] S. Smith and C. Stanfill. An analysis of the effects of data corruption on text retrieval performance. Technical report, Thinking Machines Corporation, Cambridge, MA, USA, December 1988.

[Sta81] R. Stalnaker. Probability and conditionals. In W.L. Harper, R. Stalnaker, and G. Pearce, editors, *Ifs*, The University of Western Ontario Series in Philosophy of Science, pages 107–128. D.Riedel Publishing Company, Dordrecht, Holland, 1981.

[Sus93] M. Sussna. Word sense disambiguation for free-text indexing using a massive semantic network. In *Proceedings of CIKM*, 1993.

[Sv76] K. Sparck Jones and C.J. van Rijsbergen. Information Retrieval test collections. *Journal of Documentation*, 32(1):59–75, March 1976.

[Sv93] T.M.T Sembok and C.J. van Rijsbergen. Imaging: a relevance feedback retrieval with nearest neighbour clusters. In *Proceedings of the BCS Colloquium in Information Retrieval*, pages 91–107, Glasgow, UK, March 1993.

[SY73] G. Salton and C.S. Yang. On the specification of term values in automatic indexing. *Journal of Documentation*, 29(4):351–372, 1973.

[TC90] H.R. Turtle and W.B. Croft. Inference networks for document Retrieval. In *Proceedings of ACM SIGIR*, Brussels, Belgium, September 1990.

[TC91] H.R. Turtle and W.B. Croft. Evaluation of an inference network-based retrieval model. *ACM Transactions on Information Systems*, 9(3):187–222, July 1991.

[TC92a] H.R. Turtle and W.B. Croft. A comparison of text retrieval models. *The Computer Journal*, 35(3):279–290, June 1992.

[TC92b] H.R. Turtle and W.B. Croft. Uncertainty in information retrieval systems. Unpublished paper, 1992.

[Tho89] R.H. Thompson. The design and implementation of an intelligent interface for Information Retrieval. Technical report, Computer and Information Science Department, University of Massachusetts, Amherst, MA. USA, 1989.

[Tho90a] P. Thompson. A combination of expert opinion approach to probabilistic Information Retrieval. Part 1: the conceptual model. *Information Processing and Management*, 26(3):371–382, 1990.

[Tho90b] P. Thompson. A combination of expert opinion approach to probabilistic Information Retrieval. Part2: mathematical treatment of CEO model 3. *Information Processing and Management*, 26(3):383–394, 1990.

[Tur90] H.R. Turtle. *Inference Networks for Document Retrieval.* PhD Thesis, Computer and Information Science Department, University of Massachusetts, Amherst (USA), October 1990.

[VF95] C.L. Viles and J.C. French. On the update of term weights in dynamic information retrieval systems. In *Proceedings of CIKM*, pages 167–174, Baltimore, MD, USA, November 1995.

[Voo93] E.M. Voorhees. On expanding query vectors with lexically related words. In *Proceeding of the TREC Conference*, pages 223–232, Gaithersburg, MD, USA, November 1993.

[Voo94] E.M. Voorhees. Query expansion using lexical-semantic relations. In *Proceedings of ACM SIGIR*, pages 61–69, Dublin, Ireland, July 1994.

[vR77] C.J. van Rijsbergen. A theoretical basis for the use of co-occurence data in Information Retrieval. *Journal of Documentation*, 33(2):106–119, June 1977.

[vR79] C.J. van Rijsbergen. *Information Retrieval.* Butterworths, London, second edition, 1979.

[vR86] C.J. van Rijsbergen. A non-classical logic for Information Retrieval. *The Computer Journal*, 29(6):481–485, 1986.

[vR89] C.J. van Rijsbergen. Toward a new information logic. In *Proceedings of ACM SIGIR*, pages 77–86, Cambridge, USA, June 1989.

[vR92] C.J. van Rijsbergen. Probabilistic retrieval revisited. Departimental Research Report 1992/R2, Computing Science Department, University of Glasgow, Glasgow, UK, January 1992.

[vR93] C.J. van Rijsbergen. The state of Information Retrieval: logic and information. *Computer Bulletin*, pages 18–20, February 1993.

[vRL96] C. J. van Rijsbergen and M. Lalmas. An Information Calculus for Information Retrieval. *Journal of the American Society of Information Science*, 47(5):385–398, 1996.

[Wal93] P. Wallis. Information retrieval based on paraphrase. In *Proceedings of PACLING*, 1993.

[WCS96] L. Wall, T. Christiansen, and R. Schwartz. *Programming Perl.* O'Reilly and Associates, Sebastopol, CA, USA, 2nd edition, 1996.

[WCY93] S.K.M. Wong, Y.J. Cai, and Y.Y. Yao. Computation of term association by a Neural Network. In *Proceedings of ACM SIGIR*, Pittsburgh, PA, USA, July 1993.

[Wei73] S.F. Weiss. Learning to disambiguate. *Information Storage and Retrieval*, 9:33–41, 1973.

[WY89] S.K.M. Wong and Y.Y. Yao. A probability distribution model for Information Retrieval. *Information Processing and Management*, 25(1):39–53, 1989.

[WY91] S.K.M. Wong and Y.Y. Yao. A probabilistic inference model for information retrieval. *Information Systems*, 16(3):301–321, 1991.

[WY95] S.K.M. Wong and Y.Y. Yao. On modelling information retrieval with probabilistic inference. *ACM Transactions on Information Systems*, 13(1):38–68, 1995.

[Yar92] D. Yarowsky. Word sense disambiguation using statistical models of Roget's categories trained on large corpora. In *Proceedings of the COLING Conference*, pages 454–460, August 1992.

[Zad87] L. A. Zadeh. *Fuzzy sets and Applications: Selected Papers.* Wiley, New York, 1987.

[Zer91] U. Zernik. TRAIN1 vs. TRAIN2: agging word senses in corpus. In *Proceedings of RIAO, Intelligent Text and Image Handling*, pages 567–585, 1991.