An Axiomatic Theory for Information Retrieval

Een Axiomatische Theorie voor Information Retrieval

(met een samenvatting in het Nederlands)

Proefschrift

ter verkrijging van de graad van
doctor aan de Universiteit Utrecht
op gezag van de Rector Magnificus, Prof. dr. J.A. van Ginkel,
ingevolge het besluit van het College van Decanen
in het openbaar te verdedigen
op maandag 18 november 1996 des namiddags te 12:45 uur

door

Theodorus Wilhelmus Charles Huibers

geboren op 8 januari 1966
te Valkenswaard

Promotoren: Prof. dr. J. van Leeuwen
Faculteit Wiskunde en Informatica
Prof. dr. Y. Chiaramella
CLIPS - IMAG
Université Joseph Fourier, Grenoble, France

CIP-GEGEVENS KONINKLIJKE BIBLIOTHEEK, DEN HAAG

Huibers, Theodorus Wilhelmus Charles

An Axiomatic Theory for Information Retrieval /
Theodorus Wilhelmus Charles Huibers. - Utrecht :
Universiteit Utrecht, Faculteit Wiskunde en Informatica
Proefschrift Universiteit Utrecht. - Met index, lit. opg.
- Met samenvatting in het Nederlands.
ISBN 90-393-1207-9
Omslagillustratie: Anne Besselink.

Preface

In the summer of 1992 I was finishing my masters studies with an internship post at a publishing house. It was during that period that I became attracted to the vast number of open research problems concerning information management. The retrieval of relevant information is just one problem, but one so interesting, complicated and diverse, that it proved worth writing a thesis about. In essence, this thesis focuses on one single question: how can we decide that some information is about other information. My contribution to answering that question takes about two hundred pages and I can only hope that it brings us a bit closer to a solution.

First of all, I wish to thank Peter Bruza, my masters thesis supervisor, and very soon after, my initial Ph.D. supervisor, for bringing me into the field of information retrieval and research. He introduced me to the many enjoyable aspects of academic research. It is a pity that he went back to Australia after a year, although this offered me the possibility to visit him (and Brisbane) for a month in October of 1994.

During my first year as a Ph.D. researcher, I became acquainted with the information retrieval family. One of the *family-meetings* was the annual British Computer Society Information Retrieval Colloquium. This was for me *the* opportunity to meet other young researchers and to talk, discuss, explain, learn, and become involved in the topics of the field.

At my first colloquium in 1993 I met Mounia Lalmas, who became my *"logical information retrieval"* companion. I am grateful that she was always prepared to help and support me with my research in many ways.

In September 1993, my friend Bernd van Linder, whom I knew from my masters studies at the University of Nijmegen, became a member of our department. His knowledge of research, logic, politics, football, and many other subjects is impressive and his input has had a big influence on this thesis. He was also kind enough to play the role of devil's advocate concerning situation theory, my theoretical approaches, and for many other issues (e.g. football, politics, law). During the preparation of my thesis, Bernd's feedback was invaluable. I would like to thank him for all his help, without which this thesis might never have been finished.

Another type of support, financial support, is also important for a Ph.D. researcher.

It allows him to travel and present the research in progress all around the world. For this reason, I would like to acknowledge the STIR-funding (visiting Peter Bruza in Brisbane), the HCM-funding (SIGIR'94 in Dublin), SHELL and the ICCS-committee (ICCS'96 in Sydney), and especially IDOMENEUS for allowing me to follow a research exchange program (Esprit Network of Excellence No. 6606) that gave me the possibility to spend 7 months (January 1995 till August 1995) at the LGI-IMAG lab of Yves Chiaramella in Grenoble.

It was Yves who invited me to Grenoble, to do research on logical information retrieval models. These seven months were a pleasant and productive period. In an inspiring atmosphere, I cooperated with Jean-Pierre Chevallet, Nathalie Denos, and Iadh Ounis, constantly supported by the rest of the group. As my promotor, Yves was always enthusiastic about my theoretical approach, and he kept me on the right track.

Coming back from “La douce France”, my promotor Jan van Leeuwen gave me full supervision. I am grateful to him for his patience and his interest in my work. In a cheerful way, he suggested many corrections and improvements while still maintaining my confidence in my work.

I would also like to thank Thomas Arts, Wiebe van der Hoek, and John-Jules Meyer for answering many questions, reading my drafts, and helping me, especially when I got stuck with the formal aspects of the logical framework.

In the beginning of my Ph.D. research, I came into contact with a few other young researchers in the area of Multi-Media and User Interfaces (MMuis). We decided to meet each other and present our work on an regular basis. These *MMuis* meetings were important for the pleasure in my work and I thank all the participants for turning those meetings into enjoyable and inspiring gatherings.

In 1995 I won the Dr. Ir. H.C. Molster prize for innovative research in the area of online information systems. This could not have come at a better time. It renewed my flagging determination to complete the writing process.

Working together with people with different cultural backgrounds, attitudes, interests, and ideas was what I found the most pleasant part of my work. For this reason I would like to thank the co-authors of my articles: Frits Berger, Peter Bruza, Jean-Pierre Chevallet, Nathalie Denos, Wiebe van der Hoek, Mounia Lalmas, Bernd van Linder, John-Jules Meyer, Iadh Ounis, Keith van Rijsbergen, Cees Witteveen, and Bernd Wondergem.

I want to thank the members of my reading committee: Prof. Dr Michiel Hazewinkel, Dr Bernd van Linder, Prof. Dr John-Jules Meyer, Prof. Dr Keith van Rijsbergen, and Dr Theo van der Weide, for reviewing this thesis.

Anne Besselink has drawn the beautiful cover illustration and she is gratefully acknowledged for that. Also for the contents of this thesis, I had a lot of ‘helping hands’. Besides Mounia and Bernd, I would like to thank Frits Berger, Iain Campbell, Lynda

Hardman, and Iadh Ounis for reading and commenting.

Furthermore, my thanks go to my loyal supporters Maria Wessels and Ortwin Hutten; for all the ice-creams we ate, for all the discussions we had, and for all the pep-talks.

I would like to thank my parents, who always kept me aware of my 'Brabantse' roots. As a typical example, my father still questions the benefit of a scientific researcher who does not know how to plant a potato. I still do not have an adequate answer, perhaps one day I can prove theoretically that the question is undecidable. Finally, I am very grateful to Pierre-Ine and our little Zoo. They have been the anchor of happiness in my life.

Theo Huibers, October 1996.

Contents

Chapter 1

Introduction

> *In the beginning there was information.*
> *The word came later.*
>
> F.I. Dretske, *'Knowledge and the Flow of Information'*.

In the world there is increasing competition to manage information faster and more inexpensively. This competition is driven by the explosive growth of the amount of information available. For example, every year in the European Community about two million academic, scientific, medical and social-economic articles and books appear [20].

At the other end of the spectrum is the expansion of the amount of information available via the Internet (also known as the Web). The Internet can be viewed as a connected collection of accessible information stores, the so-called hosts. The tremendous growth in the number of hosts is shown in Table 1.1[1].

Date	*Hosts*
08/81	213
05/82	235
08/83	562
10/84	1,024
10/85	1,961
02/86	2,308
11/86	5,089
12/87	28,174
07/88	33,000
10/89	159,000
10/90	313,000
10/91	617,000
10/92	1,136,000
10/93	2,056,000
10/94	3,864,000
01/95	4,852,000
07/95	6,642,000
01/96	9,472,000
07/96	12,881,000

Table 1.1: The development of the number of Internet hosts.

The digithrope Negroponte calculated that if the rate of growth of the number of Internet users were to continue at the rate at the time of writing (1996), which is of

[1] Source: http://www.nw.com/zone/WWW/top.html.

course impossible, the total number of Internet users would exceed the population of the world by 2003 [105].

To conclude, the amount of available information is currently growing at an incredible rate. Information about almost any subject is accessible leading to the presence of various information providers at the Internet and the development of powerful tools such as NCSA Mosaic and Netscape to browse, search and access this information. Since the digital highway virtually connects all parts of the world with each other, accessing information no longer presents a problem.

Be that as it may, a growing amount of information remains unused (or unread) simply because there are no means for retrieving this information effectively. In order to tackle this problem attempts have been made, since the early nineteen sixties, to create appropriate computer systems the so-called *information retrieval systems* (or: IR systems for short). In the next section we give a brief history of IR systems.

1.1 The history of information retrieval

In the early days of information storage, document collections were not large enough to be in need of an IR system, not even a manual one. For example, at the end of the fourteenth century, the English poet Geoffrey Chaucer, in addition to being famous for his poems, also became very well-known for the fact that he owned about sixty books, which had cost him an immense amount of money [111]. Of course, the number of books in the libraries was much larger, but still very modest. The library of the Sorbonne, which was known to be one of the largest in the fourteenth century, contained 1,722 books in those days [111].

With the constant growth of the number of books an overview of a collection became necessary. Such an overview was presented in a catalogue. In 1604 the Catalogue of the Bodleain, the university library of Oxford, was printed for the first time. It was the largest general catalogue of the contents of any European library published up to that time. About 10,200 titles were described in the Catalogue, including subject lists of writers on Scripture, on Aristotle, on Law and on Medicine [129].

The retrieval of documents was done by using the document references, which were first stored in catalogues and card-trays. With the arrival of the computer in the nineteen fifties, retrieval made a big step forward since as of that moment on, the references could be recorded in database systems. In the rest of this thesis we restrict ourselves to computerised systems only. Therefore, when referring to information retrieval (or IR systems), it is implicitly understood to be the computerised variant.

In the beginning of the computer era, information retrieval was focused on automatic data processing: not the information *in* a document but information *about* the document was stored. The representatives of a document were facts such as author name(s), title, number of pages, etc. These facts were recorded in database systems. The information *in* the document was only accessible if the document was in your hands. Literally, in those days information retrieval was *data retrieval*; if a requested fact "Writer: W. Shakespeare" was stored as a representation of a document entitled "Julius Caesar", then a reference (place-code) of the document "Julius Caesar" was retrieved. With this reference in hand, one had to search for the actual document placed somewhere on a shelf.

Back then it was practically impossible to store the information that was contained in a document. However, the importance of being able to store this information was already recognised, as can be seen in the following quotation taking from the book 'Automatic Data Processing' of Brooks and Iverson [21] (page 52) that was written in 1963:

> *'The future importance of any aspect of an event is, however, not easily estimated, and much data are therefore recorded and retained for potential, though unspecifiable, future use. Such retention pertains particularly to* documents, *records which because of some validation are peculiarly acceptable as evidence.'*

A few years later, the *information* retrieval systems were establishing their prominent position among the computer systems. The information content of a document was represented by keywords or other representatives. So, the goal of an IR system was changed from *data retrieval* to *information retrieval.* For instance, in 1967 Martin [96] (page 164) described the task of an IR system as follows:

> *'A real-time [IR] system may be used to provide information about a service or a situation when it is required and where it is required.'*

For instance, given an IR system, the management of a company could obtain (or may wish to obtain) information about another company; the police could do a search in their document-bases to find out whether their is a suspect fitting their descriptions; a researcher may wish to know what literature exists on her topic.

Besides the fact that in data retrieval the user retrieves descriptors and in information retrieval she retrieves the object in question, there are some other essential differences between data and information retrieval. Books about information retrieval most often start with a list of the distinguishing properties of data and information retrieval. A summary of the principle differences as given by Blair [19], Turtle & Croft [143], and Van Rijsbergen [122] is shown in Table 1.2.

Given this table, it may be inferred that one of the primary task of an IR system is to provide information, definitely not just any piece of information but information that is relevant with respect to an information need.

	Data Retrieval	*Information Retrieval*
The representation of stored information	Well-defined types of objects and facts	Unstructured information
The method of answering a request for information	Direct, through facts	Information which will likely contain what the user wants
The relation between the formulated query to the system and the satisfaction of the user	Satisfaction or no satisfaction (deterministic)	A high likelihood that the user is satisfied
The definition of a successful system	Does the system deliver the requested facts?	Does the system satisfy the users' information need?

Table 1.2: A summary of the differences between data and information retrieval.

From the nineteen seventies onward, IR systems steadily stored representatives of the information content of documents. Still, the IR systems did not provide information other than document references. For instance, Lancaster [88] formulated the terms *information retrieval* and *information retrieval system* in the context of items of literature as follows:

> *'The term information retrieval, as it is commonly used, refers to the activities involved in searching a body of literature in order to find items (i.e., documents of one kind or another) that deal with a particular subject area. An information retrieval system, then, is any tool or device that organizes a body of literature in such a way that it can be searched conveniently.'*

In the nineteen eighties, a new generation of IR systems based on Natural Language Processing (NLP) techniques was proposed. These systems dealt with the text in the document as meaningful sequences of words rather than just as character strings (see for instance [18, 136]).

In the nineteen nineties, when multi-media is no longer a futuristic but a cognitively acceptable way to represent information, the information retrieval problem starts to explode. From this moment onwards, a user is not only searching for information contained in text, but also for information contained in sounds, images or video. Fortunately, a huge part of the media is stored in digital form. Documents are no longer exclusively on a shelf but also reside on a computer's accessible storage.

Now, being able to access the information content of a document directly, the task of an IR system has increased tremendously. The provision of information rather than a reference to a document has become the primary task of an IR system. As Van Rijsbergen and Lalmas [127] recently wrote:

> *'the purpose of an information retrieval system is to provide information about a request and that a request is a representation of an information need that an IR system attempts to satisfy. [...] if a user states a query then it behoves the IR system to find the objects that contain information about that query.'*

So, from the point of view of information storage, the brief history of information retrieval started with data retrieval and ends with multi-media information retrieval.

We can also detect another line of development in the history of information retrieval. At first most IR systems were originally developed to perform the 'classical information retrieval' task: a person was looking for relevant information in *one* collection of documents (often a library). Nowadays, the information retrieval problem is much broader as a user is searching for information contained in very large information stores, possibly spread around the globe. Two typical examples are the Internet as we mentioned previously, and digital libraries. Digital libraries consider related digital documents from all over the world as belonging to their collection. The user does not notice[2] the difference between consulting a document that is physically stored[3] in Australia or in the Netherlands. The task of a digital library is to retrieve documents which contain quality[4], non-redundant[5] information about a formulated request regardless of physical storage location.

Due to this rapid shift of paradigms, new requirements for IR systems have been recognised. The vast size of present-day information domains forces users to apply distributed retrieval systems to help them search distinct areas of cyberspace. These systems will in general not be static, but can adapt to the wishes of the user. They will also have to be autonomous to a considerable extent since it will be impossible for the user to oversee and guide their behaviour in detail.

Over the past ten years, research in Artificial Intelligence has focused on defining highly autonomous systems displaying a rational behaviour and capable of solving complicated and elaborate tasks [90]. These systems, which are commonly referred to as *rational agents*, seem tailor-made for helping to solve the information retrieval problem. Rational agents, with the ability to reason, communicate, gather and maintain

[2] Or at least, the system should be able to keep this hidden for the user.

[3] In this sense, bits on a hard-disk, CD-Rom, tape or magnetic drum, and so on.

[4] Not all information is 'library' quality, such as the information in advertisement leaflets.

[5] An identical document taken from the Netherlands and Australia should not be retrieved twice.

information could probably be used as autonomous IR systems (cf. [33, 70, 72, 91]). On the Internet, information retrieval is performed by these agents: they are not called rational agents but they have fancy names such as spiders, webcrawlers, knowbots, and so on. It is commonly agreed that the success of the Web will depend strongly on the effectiveness of these agents. The following quotation from December [47] is a case in point:

> *'In an increasingly thin soup of redundant, poor quality, or incorrect information, even the smartest Web spiders won't be very effective. A flood of information unfiltered by the critical and noise-reducing influences of collaboration and peer review can overwhelm users and obscure the value of the Web itself. The Web certainly needs solutions in information discovery and retrieval –indeed, developing intelligent spiders, worms, robots, and ants is crucial to making sense of the Web.'*

One way of ensuring these changes is to combine and improve existing IR systems. This can only be done if we have a deep insight in the retrieval process, based on what is needed and what we already have. However, in the past thirty years of information retrieval research it has become clear that it is not evident at all how to analyse, compare, and improve the retrieval processes of different IR systems. Our main aim is to define a framework that allows us to model IR systems in order to gain a better insight into the retrieval process.

1.2 Information retrieval paradigms

We begin with the old information retrieval paradigm to introduce the core of the concepts of information retrieval used in this thesis. The notation we use is given in brackets.

Information retrieval begins with a user having an information need (N) that she wishes to fulfil. The information need is formulated, as well as possible, in the form of a query (q). Often this query is constructed using a query language in the context of an appropriate user interface. Results are documents or parts of documents[6] that suit the user's need according to the IR system. The retrieved documents are taken from the document-base ($\mathcal{D}$). Normally, an IR system does not, or cannot, incorporate the entire information content of a document on account of factors of efficiency and complexity. Therefore, an IR system handles a manipulable representation of the document's

[6] A part of a document can be viewed as a (sub-)document, therefore we omit the phrase 'or parts of the document' in the rest of this thesis when we speak of documents. However, the reader should keep in mind that the ability to retrieve parts of a document is important. In our view information retrieval should not consider only a document as an atomic entity.

information content. The determination of the document representation, which is an approximate representation of the documents content, is arrived at by a process termed indexing. The indexing process returns for each document (d) a descriptor set $(\chi(d))$ as its representation.

The heart of the IR system consists of a matching process that compares the request q with the descriptor set $\chi(d)$ of each document d. If the matching operation deems a document as satisfying to the request, the document is assumed *relevant*. The matching operation is also termed the *relevance decision*. The documents that are relevant according to the system, are retrieved and displayed to the user. Very often these documents are ordered according to a degree of relevance, known as the *ranking* of the documents. In most current IR systems, the user is able, supported by the system, to reformulate the query after inspecting the result. This process is termed *relevance feedback*.

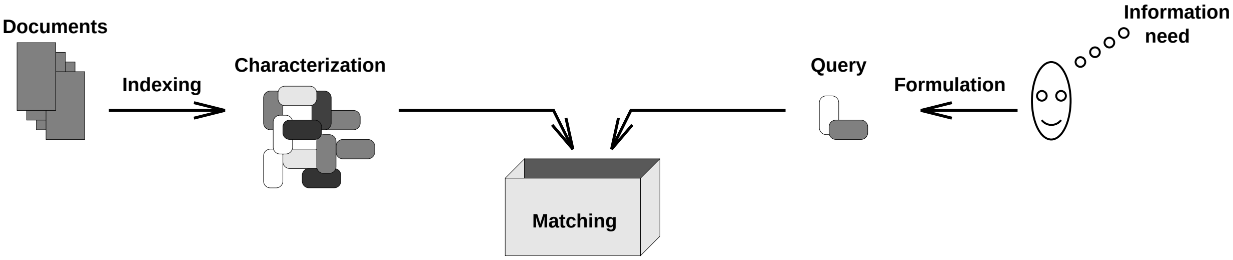


Figure 1.1: Old information retrieval paradigm.

In general, IR systems are developed from a predefined *information retrieval model* (or: IR model for short). Such a model tries to furnish an answer for the relevance decision. It will explain the structure and processes of these systems, and clarify their general, as opposed to specific, characteristics [142].

A fairly diverse range of IR systems has been proposed during the past thirty years. To date, there are boolean retrieval systems, coordinate retrieval systems, vector-space retrieval systems, probabilistic system, logical systems, etc. In Chapter 4 we present these models in more detail. For a presentation of about a score of IR systems we refer the reader to the proceedings of SIGIR [22, 53, 80]. The diversity of the IR systems results from the many possible perspectives that can be selected from the range of the relevance decisions.

As mentioned in Section 1.1, the information retrieval problem has become much broader. An information retrieval paradigm should not consist of a single matching process of one particular IR system but a collection of autonomous IR systems, which are able to cooperate amongst each other. Therefore, we present a new information retrieval paradigm to capture these new concepts.

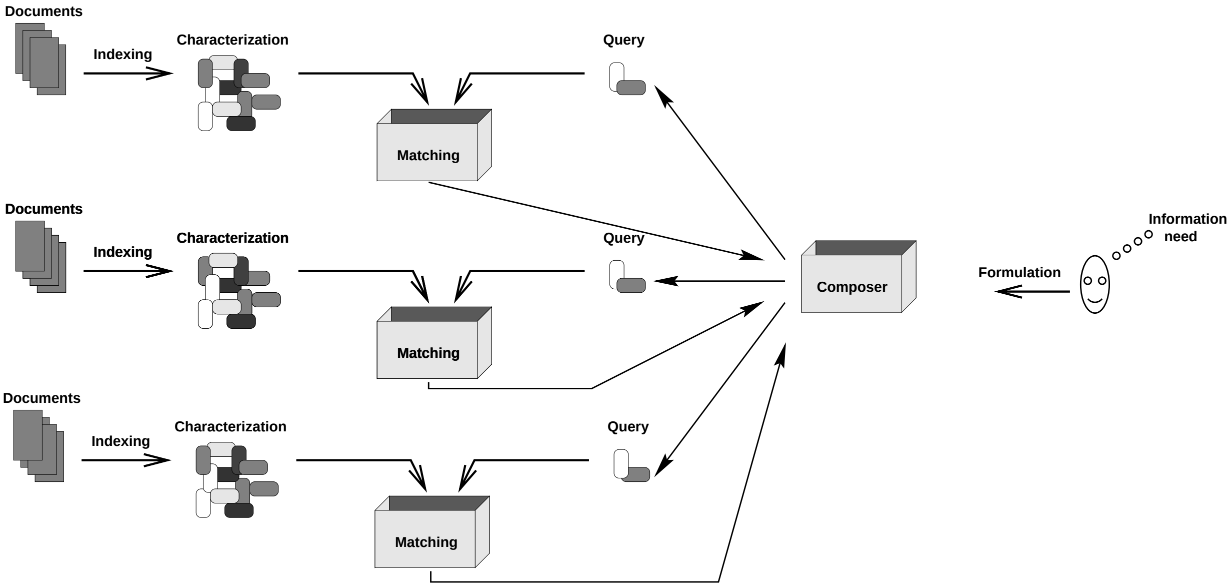


Figure 1.2: New information retrieval paradigm.

In the new information retrieval paradigm the concept of the documents remains the same. A user formulates her information need as a query (q). A *composer* translates this query into acceptable forms for a variety of IR systems. There is no single collection of documents anymore but there are numerous information stores. Each IR system returns relevant information to the composer module that filters and ranks the delivered information. The final output will be a list of all relevant information that is presented to the user.

1.3 What this thesis is about

As one can see in the paradigms, information retrieval concerns the problem of retrieving from a given document-base those documents that are *likely* to be *relevant* to a certain information need. Hence *relevance* is essentially a relation between a document and an information need. Due to the intrinsic vagueness of terms such as 'information need' and 'likely', the notion of relevance is hard to formalise mathematically.

In 1971, Cooper introduced an objective part of relevance termed *logical relevance* [41]. This logical relevance is one of the constituents of the definition of relevance. Cooper distinguishes two aspects of the notion of relevance:

- ▷ **Utility**, which describes the ultimate usefulness of the retrieved document, and
- ▷ **Logical Relevance**, which describes whether a retrieved document has some topical bearing on the information need in question.

Of course a document can be logically relevant for a given information need but at the same moment not be useful at all, for instance, because the content of the document is out of date. In this thesis we focus on the concept of logical relevance, which is described by Cooper [41] as follows:

> '*A stored sentence is* logically relevant *to a representation of an information need if and only if it is a member of some minimal premise set of stored sentences for some component statement of that need.*'

Logical models in information retrieval attempt to encompass this definition in the core notion of relevance. Following Maron [95], we call the logical relevance relation *'aboutness'*. In the literature several other descriptions of the aboutness relation can be found:

- *'topically related'* [41],
- *'correspondent to'* [107], and
- *'likely to contain information about'* [125].

The explosive growth of information has made it a matter of survival for companies, Internet users, librarians and indeed anyone dealing with information, to have *good* IR systems at their disposal. For instance, in the previous section we mentioned that the usefulness of the Web depends on how well an information retrieval agent works. In this thesis we argue that the concept of *goodness* in the context of information retrieval is related to the characteristics of the aboutness decision. In this thesis we present a framework that allows one to postulate these characteristics. With these postulates in hand it should be possible to decide that one IR system is better than another. In our opinion, the critical analysis of IR systems can be made on the basis of the aboutness decision as proposed by their underlying model.

To develop a new improved IR model a deep understanding of the relevance decisions of the various existing models is needed. In fact, we think it is more important that we develop a general information retrieval theory that offers the opportunity to define relevance independent of the IR models than to define a new IR model as such.

A general information retrieval theory should focus on the modelling of the information retrieval concepts and especially on the retrieval functions. The theory should abstract from specific notions and practical implementation problems. Therefore, such a theory is called a *meta-theory*. Above all, a meta-theory should extend the possibilities of the comparison of IR systems, which are now typically experimental. For then it becomes possible to compare IR models of different IR systems. Summarising, the study of the logical relevance definition in terms of an information retrieval theory of various IR models is the leitmotif of this thesis.

1.4 Outline of this thesis

In this thesis we propose a general framework for use in information retrieval. This general framework should provide a deeper insight in the notion of relevance used in information retrieval. It is important to have such a formalism in order to propose, build or merge IR systems that can manage any amount and sort of information as it occurs in the new information retrieval paradigm.

In Chapter 2 we present the reasons for our choice for a theoretical approach rather than an experimental one. An information retrieval theory can be split into two main parts: the representatives of information and the relevance decision. In Chapter 3, we therefore propose Situation Theory as the language for information representatives, and an aboutness proof system for the relevance decisions. With this in hand we investigate in Chapter 4 the underlying logic of relevance of different existing IR models. The objective of Chapter 4 is also to filter out suitable new rules for relevance decisions. In Chapter 5, we show that the relevance decisions underlying various IR models in the proposed theory can indeed be compared formally rather than experimentally. This means that theorems can be proved stating, for example, that the systems based on a vector-space model have exactly the same relevance decision as those based on probabilistic models. This result does not only spare us the effort of experimentation, but, more importantly, it allows us to side-step the controversies surrounding the experimental process. Thus, a classification based on the relevance decision can be made. In order to build sophisticated IR systems capable of performing the IR tasks as presented in the Section 1.2, we need to do more than a simple comparison. All these advantages of our theory are tested in Chapter 6. In this chapter we show the possibility of a ranked output by ordering a set of IR systems on a qualitative basis. We show that information retrieval agents can be modelled using the description of relevance. Furthermore, an extended example is presented, based on a specific search strategy in a hypermedia model. Finally, Chapter 7 presents the main results of the studies. In addition, conclusions are drawn with respect to the theoretical approach. We will also give recommendations for further research on the theory and its application.

Chapter 2

Approaches for studying information retrieval

Information retrieval researchers are like automotive engineers who are trying to improve the design of automobiles without being able to measure horsepower or fuel efficiency.

D.C. Blair, *'Language and Representation in Information Retrieval'*.

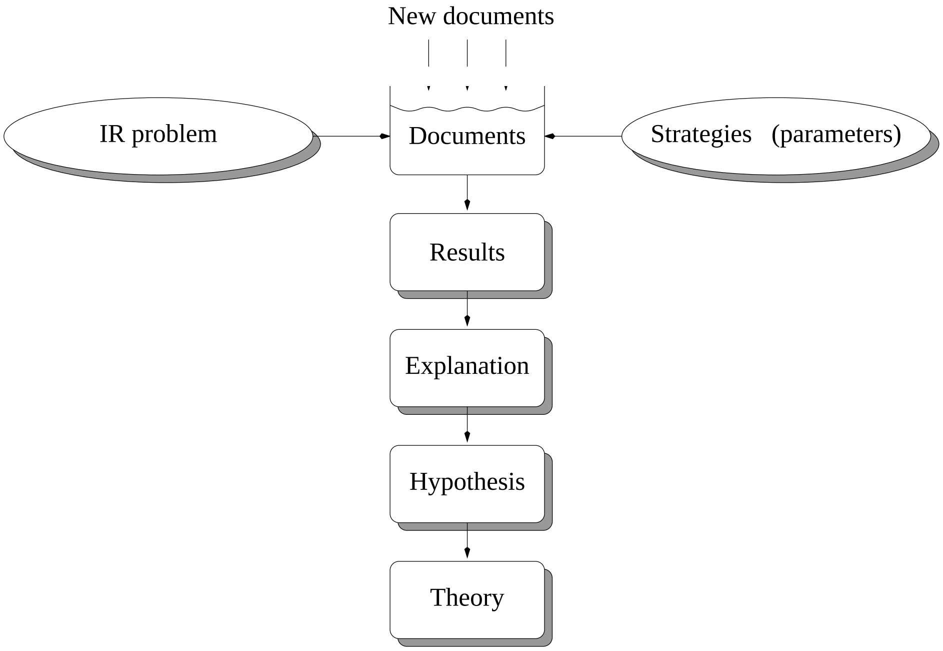
In this chapter we present two different possibilities for studying information retrieval. According to several authors [19, 43, 132] there are two possible avenues to follow for an information retrieval study, namely an experimental one and a theoretical one. The two approaches appear difficult to reconcile. The controversy between the people who were mainly inspired by mathematics and those who were influenced more by the empirical sciences originated in the field of analytical philosophy and dates back to the time of Pythagoras [130]. The article 'The Formalism of Probability Theory in IR: A Foundation or an Encumbrance' by Cooper [43] is devoted to the internal struggle between these two strategies for information retrieval. Cooper started his article as follows:

> *'Some approaches to retrieval system design are strongly guided by theory. Others have little real theoretical underpinning, but are instead more experimental and ad hoc in character. Which is preferable? Obviously, theory-guidedness is a good thing if the theory leads to promising retrieval rules to try out. Good theories have inferential power, and inferential power can help minimize empirical floundering. However, having to stay within the constraints of a strict theoretical formalism can also impose costs and penalties. The true extent of these costs is not always fully recognized.'*

This thesis proposes a theoretical formalism for information retrieval. However, before presenting the theory, we feel that it is first necessary to motivate the choice for a theoretical rather than an experimental approach. We highlight two famous information retrieval experiments, and discuss some of their observed limitations. We then present several current theoretical approaches for information retrieval. In this chapter we present two, at first sight, controversial approaches. In Section 2.2.4 we show a technique for *'theory performance'* inspired by the work of the philosopher Popper. He proposes a synthesis of both the theoretical and experimental approaches in which theories can be compared using test statements obtained from experiments. In the final section we introduce our version of the theoretical approach, tailor-made for information retrieval.

2.1 Experimental approach

Traditionally the study of IR systems is purely experimental in the sense that it is 'based on tests one planned in order to provide evidence for or against a hypothesis'[1]. The experimental study is based on the paradigm as depicted in the figure below:



In an experimental approach an arbitrary information retrieval problem would be studied as follows. An information retrieval tool is proposed which could offer a solution for a typical information retrieval question. For instance: 'is it true that adding more

[1] Another interpretation, as for instance mentioned by Van Rijsbergen, is that experimental information retrieval is mainly carried out in a 'laboratory' situation [122].

documents to the document-base automatically leads to a higher recall in the context of this system'. Varying the setting of the parameters of the system and the number of documents, a collection of results is obtained. The results are studied and could possibly form an answer to the question under scrutiny. If such an answer is validated for several systems or in various settings, a hypothesis is formulated. When the hypothesis fits in a series of other related hypotheses, a theory is proposed.

In an experimental approach the comparison of two systems is formulated as: 'is system A better than system B with respect to C' and proceeds according to the method described above. A widely accepted manner of comparing experimental results of the retrieval performance of two IR systems is the study of the so-called *recall* and *precision* measures. These measures are used to conclude whether system A is to be preferred over system B or vice versa.

2.1.1 Recall and precision

Recall and precision are two measures with as input the following two document sets: (1) the collection of documents which the user would judge to be relevant with respect to her information need, if she would be aware of all available documents (denoted as Rel_{user}), and (2) the collection of documents which an IR system retrieves according to the formulated query (denoted as Ret_{system}). The utopia of every IR system developer is to create an IR system such that the set Rel_{user} is identical to the set Ret_{system}.

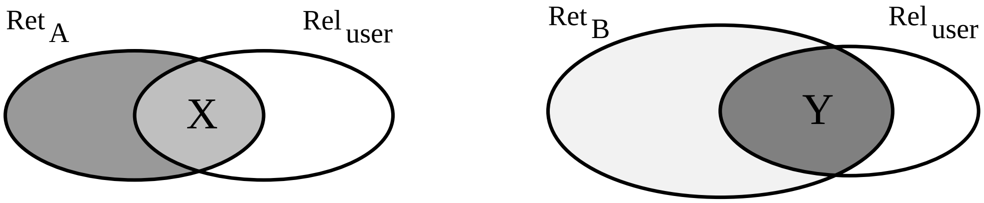
Up to now, IR systems are compared in effectiveness through a calculation using the recall and precision values. Formally the measures are defined as follows:

Definition 2.1

$$Recall = \frac{\mid Rel_{\text{user}} \cap Ret_{\text{system}} \mid}{\mid Rel_{\text{user}} \mid} \qquad Precision = \frac{\mid Rel_{\text{user}} \cap Ret_{\text{system}} \mid}{\mid Ret_{\text{system}} \mid}$$

Note that both recall and precision values are between 0 and 1. A high recall value implies that most of the documents that are deemed to be relevant according to the user are actually returned by the system. A high precision value indicates that most of the documents that are returned by the system are indeed relevant in the perspective of the user. In probabilistic terms one can rewrite the values as: $Recall = Pr(Ret_{\text{system}} \mid Rel_{\text{user}})$ and $Precision = Pr(Rel_{\text{user}} \mid Ret_{\text{system}})$.

It is now generally recognised that there is usually a certain trade-off between both values: one could naively try to increase the recall value of a system by increasing the number of returned documents, but in general this will lead to a decrease of the precision value [30]. Let us digress briefly and provide one way to explain this phenomenon in terms of *'incremental precision'*. In the following figure the results of two IR systems A and B are graphically depicted.



Let us assume that system B retrieves more documents than system A (on a query) in an attempt to achieve a better recall. Thus let $\mid Ret_{\mathsf{B}} \mid > \mid Ret_{\mathsf{A}} \mid$. By abuse of notation we will write $\#A = \mid Ret_{\mathsf{A}} \mid$ and $\#B = \mid Ret_{\mathsf{B}} \mid$. We then have:

Lemma 2.1

$$Prec_A > Prec_B \Leftrightarrow Prec_A > \frac{\mid Rel_{\mathsf{user}} \mid \times (Recall_B \Leftrightarrow Recall_A)}{(\#B \Leftrightarrow \#A)}.$$

Proof Observe that

$$\begin{aligned}
Prec_A > Prec_B \quad &\Leftrightarrow \quad \#B \times Prec_A > \#B \times Prec_B \\
&\Leftrightarrow \quad (\#B \times Prec_A \Leftrightarrow \#A \times Prec_A) > (\#B \times Prec_B \Leftrightarrow \#A \times Prec_A) \\
&\Leftrightarrow \quad (\#B \Leftrightarrow \#A) \times Prec_A > (\#B \times Prec_B) \Leftrightarrow (\#A \times Prec_A) \\
&\Leftrightarrow \quad Prec_A > \frac{(\#B \times Prec_B) \Leftrightarrow (\#A \times Prec_A)}{(\#B \Leftrightarrow \#A)}.
\end{aligned}$$

Let $X = Ret_{\mathsf{A}} \cap Rel_{\mathsf{user}}$ and $Y = Ret_{\mathsf{B}} \cap Rel_{\mathsf{user}}$. Note: $Prec_B = \frac{\#Y}{\#B}$ and $Prec_A = \frac{\#X}{\#A}$, thus

$$\begin{aligned}
Prec_A > Prec_B \quad &\Leftrightarrow \quad Prec_A > \frac{(\#Y \Leftrightarrow \#X)}{(\#B \Leftrightarrow \#A)} \\
&\Leftrightarrow \quad Prec_A > \frac{(Recall_B \times \mid Rel_{\mathsf{user}} \mid) \Leftrightarrow (Recall_A \times \mid Rel_{\mathsf{user}} \mid)}{(\#B \Leftrightarrow \#A)} \\
&\Leftrightarrow \quad Prec_A > \frac{\mid Rel_{\mathsf{user}} \mid \times (Recall_B \Leftrightarrow Recall_A)}{(\#B \Leftrightarrow \#A)}.
\end{aligned}$$

□

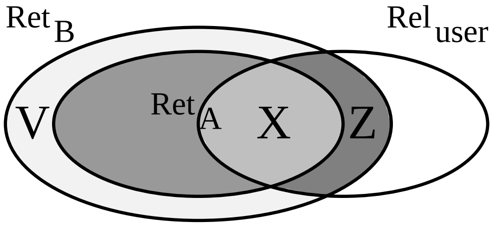
Thus if we retrieve more documents but recall goes down ($Recall_B < Recall_A$), also precision goes down. But if recall goes up ($Recall_B > Recall_A$), we have a strict condition for precision to go up or not.

Consider the common case in which system B returns at least all those documents that are returned by system A and possibly more, i.e., $Ret_{\mathsf{A}} \subseteq Ret_{\mathsf{B}}$ and let us see when recall goes up. Let $\Delta_{B,A} =^{def} Ret_{\mathsf{B}} \Leftrightarrow Ret_{\mathsf{A}}$.

Definition 2.2 The *incremental precision* of B over A is

$$IP_{B,A} =^{def} \frac{\mid \Delta_{B,A} \cap Rel_{\mathsf{user}} \mid}{\mid \Delta_{B,A} \mid}.$$

Under the assumption that $Ret_{\mathrm{A}} \subseteq Ret_{\mathrm{B}}$, we have the following figure.

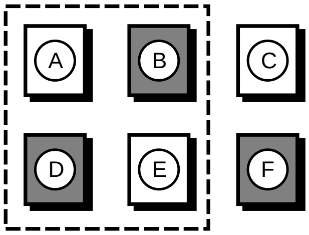


Here, Z represents the set of retrieved documents of B that are not already in Ret_{A} and that are relevant according to the user. The set V represents the retrieved documents of B that are not yet in Ret_{A} and that are not relevant according to the user. Thus, $\Delta_{B,A} = V \cup Z$. According to the previous inequality, applying Lemma 2.1 precision goes down in this case if and only if

$$\begin{aligned} Prec_A > Prec_B \quad &\Leftrightarrow \quad Prec_A > \frac{(\mid Z \mid + \mid X \mid) \Leftrightarrow \mid X \mid}{\mid Z \mid + \mid V \mid} \\ &\Leftrightarrow \quad Prec_A > \frac{\mid Z \mid}{\mid Z \mid + \mid V \mid} = IP_{B,A}. \end{aligned}$$

It follows that if $Recall_B$ is greater than $Recall_A$ (thus Z is not empty), then the precision has increased if and only if $Prec_A < IP_{B,A}$. In other words: if and only if B achieves a better precision among the extra documents in Z and V than A did on the original set of returned documents (which B returns to by assumption). This is indeed unlikely, explaining the observed phenomenon to some extent.

Example 2.1 The figure below shows a retrieval situation. The grey blocks represent the documents belonging to the set judged relevant by the user (the set Rel_{user}). The white blocks represent irrelevant documents. The blocks in the large square were retrieved by a certain IR system A (the set Ret_{A}).



Since $\mid Rel_{\mathrm{user}} \mid = 3$, $\mid Ret_{\mathrm{A}} \mid = 4$ and $\mid Rel_{\mathrm{user}} \cap Ret_{\mathrm{A}} \mid = 2$, it follows that the recall value is $\frac{2}{3}$ and the precision value is $\frac{2}{4}$. In order to achieve a better recall and a better precision with a system B, the notion of incremental precision requires that $IP_{B,A} > \frac{1}{2}$. Thus, the precision increases if and only if $Ret_{\mathrm{B}} = \{A, B, D, E, F\}$, because then $IP_{B,A} =$

1 and $Precision_B = \frac{3}{5}$. In all other cases $IP_{B,A} \leq \frac{1}{2}$, i.e., if $Ret_{\mathsf{B}} = \{A, B, C, D, E\}$ then $IP_{B,A} = 0$ and if $Ret_{\mathsf{B}} = \{A, B, C, D, E, F\}$ then $IP_{B,A} = \frac{1}{2}$. For the last case, the precision will neither increase nor decrease.

As we will show in Chapter 5, one of the advantages of our theoretical approach is that it allows predictions to be made about the recall-value with respect to the underlying IR model and information domain.

The main problem of recall and precision measures is the determination of the set Rel_{user}. In the next section we present two information retrieval tests. These tests show the construction of such sets. Important for the success of the evaluation is the ability to construct a suitable test collection. According to Hull [75], the fundamental components for a successful evaluation of a retrieval experiment are the availability of the following:

1. At least one document collection suitable for testing. The collection must include a number of queries and their relevance assessments. The relevance assessments determine sets of documents which are relevant, given the query (Rel_{user});
2. A measure, based on the similarity ranking of relevant and non-relevant documents with respect to the query, that reflects the quality of the search; and
3. A valid (statistical) methodology for judging whether measured differences between retrieval methods can be considered statistically significant.

Next we present two of the more common information retrieval tests to give the reader an idea of how these tests are brought about.

2.1.2 Cranfield

One of the very first well-known information retrieval tests was the Cranfield test[2]. The test took place in the nineteen sixties. Strictly speaking, there were two Cranfield tests, both of which we present briefly.

Cranfield 1 The goal of the first Cranfield test was a comparative evaluation of four IR systems. It was a two-year project and the evaluation was focused on the indexing process rather than on the matching function. Four different indexing processes were examined:

1. Conventional Classification,
2. Alphabetical subject index,
3. Devised schedule of a facet classification,
4. Uniterm System of Coordinate Indexing.

[2] We recommend Cleverdon's article [37] for a detailed overview of the Cranfield tests.

The document collection consisted of 18,000 papers on aeronautical engineering. In addition, 1,200 queries based on a single document in the test collection were created (in our terminology, for each query Rel_{user} was a singleton set with the document based on the query in it). A search was successful if Ret_{system} contained the document used to create the query. The results showed that all four systems were 74 - 82% effective in retrieving the required document. The analysis was based on aspects such as time for indexing, learning process, and number of returns. The major point of criticism from the information retrieval community was that the construction of the search questions was based on documents in the test collection. So, in this case one cannot speak of a query which is formulated by someone with an information need.

Cranfield 2 The second Cranfield test kept the focus on the indexing process. The objective was to examine the effect of index languages, in isolation or in any possible combination, using recall and precision measures. The document collection was created in a way totally different from Cranfield 1.

Two hundred authors of recently published papers were asked to state in the form of a question the problem which their paper addressed. Furthermore, they had to add supplementary questions that arose in the course of their research. They were then requested to indicate, on a scale of 1 to 5, the level of relevance to each question of the references they had cited in their paper. Out of these references 1400 documents were selected and 279 queries were inspected by students and by the originator of the question.

The evaluation concentrated on the indexing of the documents. Here, the tests used a multi-stage process of indexing. An indexer manually identified the concepts in the document with one, two or three keywords. A weight in the range of 1 to 3 was assigned to each concept according to the importance for a particular document. Each single word was then listed with respect to the values of the concepts it occurred in. Finally the concepts were combined into themes. Given this indexing process, different representation languages were studied. Figure 2.1 is taken from Cleverdon [37] and presents the way the languages were obtained.

For each question it was inspected which documents would be retrieved and at which level (the latter aspect is important for the recall and precision ratio). The results were presented in the form of recall- and precision-curves.

Each question was indexed in all different representation languages. The results of the Cranfield 2 tests were not as evident as those from the Cranfield 1 tests. Salton [132] indicated for instance, that ‘it is also the first evaluation project that produced unexpected and potentially disturbing results.’ Among others, Salton was surprised by the result that an advanced indexing process (concept indexing) showed a worse performance than the simple indexing process (keywords indexing).

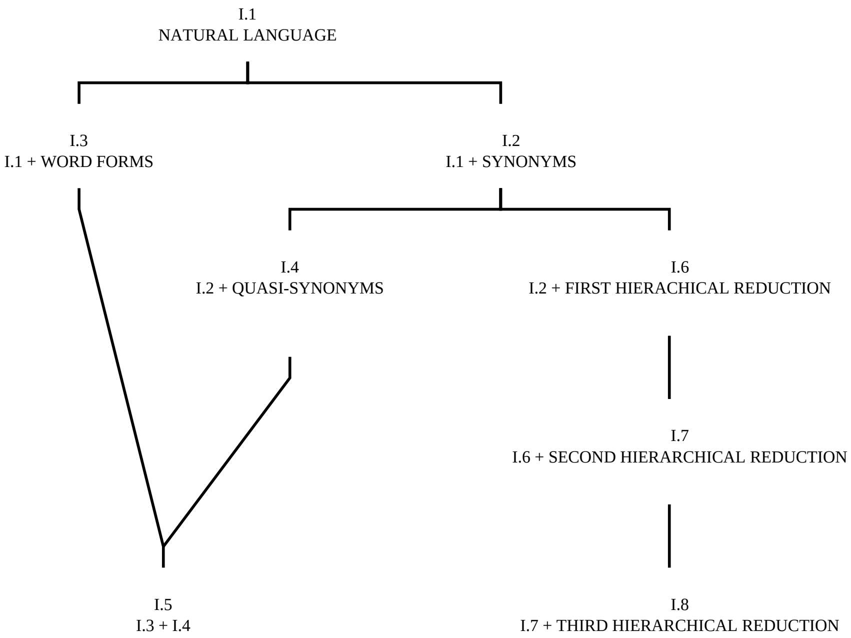

I.1
NATURAL LANGUAGE
I.3
I.1 + WORD FORMS
I.2
I.1 + SYNONYMS
I.4
I.2 + QUASI-SYNONYMS
I.6
I.2 + FIRST HIERACHICAL REDUCTION
I.7
I.6 + SECOND HIERARCHICAL REDUCTION
I.5
I.3 + I.4
I.8
I.7 + THIRD HIERARCHICAL REDUCTION

Figure 2.1: Cleverdon's representation languages.

After such a result there are two possible avenues: show that the test is not applicable and therefore the results are meaningless, or change the system in such a way that it will perform better with respect to the test. With respect to the former, it was remarked that the relevance assessments with respect to the corresponding queries might benefit the simple matching techniques at the expense of the more complex matching techniques. In Section 2.2.4 we will return to this aspect of model performance.

2.1.3 TREC

In November 1992 the first TREC was held. TREC is the acronym of the annual Text REtrieval Conference. Its proceedings [60, 62, 63] contain papers about tests and their results. The tests are elements of the TREC programme, an officially organised activity, which has as its main goal to study different approaches to the retrieval of text for large document collections. At the moment, TREC is the major experimental effort in the information retrieval field[3]. The test collection contains approximately one million documents (about 3 gigabytes of data). To compare the results obtained there is a detailed schedule that all the participants of TREC should obey. For TREC-4 the schedule was as follows:

[3] We recommend 'Overview of the Third Text REtrieval Conference (TREC-3)' [61] by Harman and 'Reflections on TREC' by Sparck Jones [139] for a detailed overview of TREC.

Jan All potential participants should apply for a position in the tests. The program committee looks for as wide a range of text retrieval approaches as possible, and selects only those participants who are able to work with the large data collection.

Feb For the accepted participants there are 3 gigabytes of information with queries and relevance judgements available (taken from previous TREC tests). With this collection they are able to train and improve the performance of their system.

May A list of routing topics is distributed.

June The test data is sent to the participants.

July Fifty new test topics for ad hoc test are distributed.

Aug The results should be submitted.

Oct The evaluation process takes place.

Nov The obtained results are presented during the TREC conference.

As one can see, there is only one month of time to process the queries. The very large amount of information makes it almost impossible to have manual interference in the indexing or matching process. Another point of interest is the distinction made between *routing* and *ad hoc* test topics. In the routing test mode, the situation is simulated in which the same questions are always being asked but new or more data is being searched. This task is similar to the one done by news clipping services or by library profiling systems [61]. Then, the relevance decision depends on previous results, and thus a kind of *learning process* is involved. In an ad hoc test mode the document collection is fixed and the question is variable. The question 'is this document relevant for the query' is considered independently of previous results. This task is similar to how a researcher might use a library, where the collection is known but the questions which are likely to be asked are not [61]. Therefore the time for processing ad hoc tests is much shorter than for the routing ones.

The document set in TREC is taken from several sources, individually varying and collectively varying in topics and genres, though with much news story material. While the relevance judgements in the Cranfield tests were done manually for the complete collection, for the TREC-collections this was practically impossible (3 gigabytes!). All TRECs have used the pooling method [140], which proceeds as follows: for each query and for each system the top 100 retrieved documents are merged in a pool[4], which is then shown to human assessors.

According to Harman [61], an important underlying assumption of this retrieval test is 'that the vast majority of relevant documents have been found and that documents that have not been judged can be assumed not to be relevant'.

[4] For TREC-1 and TREC-2, for TREC-3 it was 200.

The main results of TREC indicated not so much that one technique was shown to be significantly better than another one, but rather that individual retrieval systems were improving over time. Also, the achieved evaluation performance is an important result of TREC.

2.1.4 Reflections on the experimental approach

Since the early nineteen sixties experimental retrieval evaluations have been constructed. Very often the results of these evaluations were criticised by the experts (for an overview of various arguments see [132]). The next quotation from Cleverdon [37], one of the originators of the Cranfield tests, makes this particularly clear:

> *'The publication of the final report [36] attracted wide interest, caused considerable annoyance to the advocates of the different systems, and received some praise but much criticism.'*

Obviously some criticism was justified. A whole new discipline in the area of information retrieval developed, i.e., the evaluation of information retrieval evaluations. One typical example of such a meta-evaluation question is the following given by Hull [75]:

> *'Why should experimental results based on collections with a very limited number of short documents on restricted topics be applicable to much larger and more variable documents collections that are found in real retrieval settings?'*

This intuitively acceptable concern was the underlying reason for the TREC-community to ensure that the amount of information in the test collection was enlarged.

The following list presents an overview of the main 'concerns' made by evaluation analysts [19, 40, 75, 132]:

(i) The current measures, such as recall and precision, are not properly representing the *acuity* of an IR system because

- ▷ there is a retrieved and a unretrieved set of documents, without taking into account the possibility of an order of retrieval involving more than two classes;
- ▷ the utility factor of a document is not measured;
- ▷ returning a larger number of relevant documents is not always better: it may be that if the system highlights one relevant document this could be much more informative than returning a whole set;
- ▷ most often the measures are not dealing with interactive IR systems (such as relevance feedback systems).

(ii) The relevance assessments are not realistic, as the assessments are based on a formulated query rather than on an information need;

(iii) The evaluation must be based on knowledge of the complete set of relevant documents with respect to each query. This is hardly possible given the very large test collection;

(iv) Small document collections are not representative for large document collections (and vice versa).

To conclude, up to the time of writing IR systems are compared using statistical values such as recall and precision. We certainly perceive the utility of statistic values. However, to be able to make more strict statements concerning the qualities of one IR model when compared to another IR model, we feel that we should have more formal means of comparison at our disposal. Furthermore, to prove specific statements concerning the behaviour of IR systems, statistical tests are not adequate. There seems to be a definite need for a more formal characterisation of IR systems. In the following section we inspect theoretical approaches.

2.2 Theoretical approaches

One of the explanations of the word 'theory' in the Collins dictionary is *'a plan formulated in the mind only'*. This is certainly not what we have in mind. We prefer another description given in the dictionary, namely *'a set of hypotheses related by logical or mathematical arguments to explain a wide variety of connected phenomena in general terms'*. In this thesis the connected phenomena refer to the various stages of the retrieval process. There are various ways to study information retrieval in a theoretical way. We distinguish three approaches, namely

1. *Embedding*, which formalises an IR model that covers several other models.
2. *Categorisation*, which classifies different IR models based on a list of properties.
3. *Meta-theory*, in which a formalisation of a model and its properties in terms of a theory are presented.

These approaches do not necessarily exclude each other. One can propose a new model in terms of a theory and show how other models can be embedded. Or, one can describe properties in a theory and categorise existing IR models as to how they fulfil the described properties. In order to give the reader the essence of each approach, we discuss each of them briefly.

2.2.1 Embedding

In the case of embedding, different models are studied by mapping them to one model. We give one typical example, namely the *Inference Networks* as proposed by Turtle & Croft [143].

In the approach of Turtle & Croft different models are studied by mapping them to so-called *inference networks*. In an inference network retrieval model, retrieval is viewed as an 'evidential reasoning process in which multiple sources of evidence about document and query content are combined to estimate the probability that a given document matches a query' [143].

In an inference network retrieval model there are two directed, acyclic dependency graphs (networks), that are connected with each other.

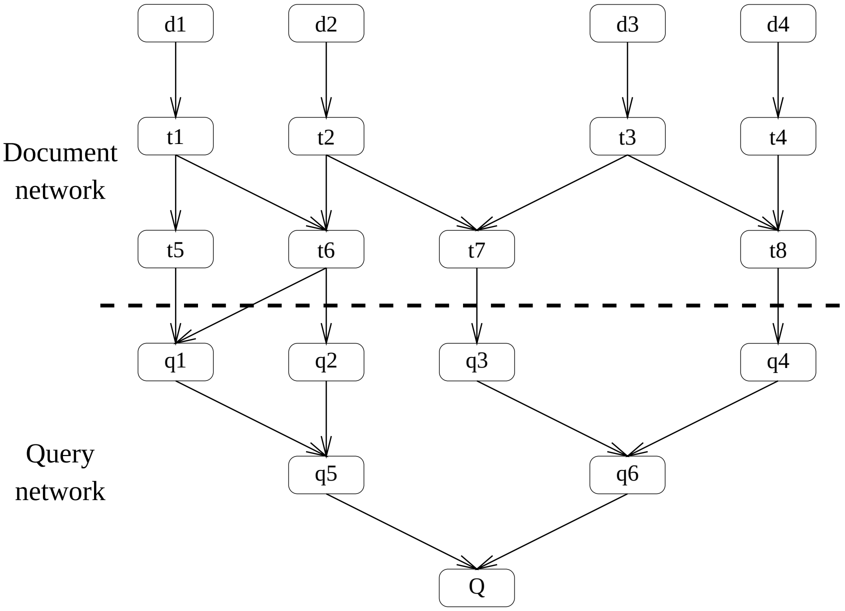


Figure 2.2: Basic inference network.

There is one graph for the document-base representation and one for the query. In the document network there are document nodes corresponding to abstract documents, text representation nodes representing information items of the document, and concept presentation nodes representing type information of the objects; the arrows in the document network represent the dependency relations. The query network is an 'inverted' directed acyclic dependency graph with a single leaf that corresponds to the event that an information need is met, and multiple roots that correspond to the concepts that express the information need. The query concept nodes define the mapping between the concepts used to represent the document collection and the concepts that make up the queries (the dotted line in Figure 2.2, shows where the mapping takes place). In the simplest case, the query concepts are constrained to be the same as the representation concepts, and each query concept has exactly one parent representation node (for instance, in Figure 2.2 the query node q_3 has as parent node t_7). In a more advanced network, the query concept may have more representation nodes (as depicted in Figure 2.2 where q_1 has as representation nodes t_5 and t_6).

By representing known models such as the boolean model, the vector-space model and the probabilistic model in terms of this network model, the authors showed that 'differences between current-generation retrieval models can be explained as different ways of estimating probabilities in the inference network' [143]. By tuning or adjusting the network they are aiming to achieve the best retrieval performance. These results can then be used for proposing a *new* IR system. For instance, the INQUERY retrieval system [31, 32] is a system based on the network model using the results from the investigation of several different models.

2.2.2 Categorisation

One typical example of an evaluating process based on categorisation is the work of Blair [19]. Blair proposes that, in order to improve information retrieval, a good theory of document representation is needed which is primarily based on language and meaning. In order to study different approaches for modelling information in information retrieval, twelve principal formal models are defined. For example Blair's 'Model 12' is presented as follows:

Example 2.2 Model 12 (Weighted Thesaurus)

I. Requests are single terms.

II. Index assignments: a set of one or more descriptors.

III. Documents are either retrieved or not.

IV. Retrieval rule: the request descriptor is looked up in a thesaurus (on-line) and semantically related descriptors above a given cut-off value (weight) are added (disjunctively) to the request descriptor. The cut-off value could be given by the inquirer.

After presenting a model, advantages and disadvantages are summed up. For instance, one of the advantages of model 12 is that it provides the user with a list of terms which are semantically related to those in the thesaurus, which is especially useful in systems with uncontrolled vocabularies [19].

As mentioned in Section 2.1.4, Blair is one of the critics of an evaluation based on statistical estimations only. The great expense of such evaluations (in time and cost) may prevent them from being performed very often (such as the scheduling of TREC, which covers almost a whole year!). Blair compares information retrieval to astronomy and quantum physics where experiments are expensive but a theoretical formalism exists that can be used to advance theoretical understanding of these disciplines independently from empirical verification. Blair states:

> *'Information retrieval would benefit greatly from the development of a similar theoretical formalism that would permit at least some of its advances to be done independently from empirical validation.'*

His book [19] is focused on the use of language and its representation. Our approach is more directed towards the relevance decision. However we believe that both approaches could be used in tandem in order to study IR models.

2.2.3 Meta-theory

In a meta-theoretical approach, information retrieval is viewed in terms of a theory T. The model and its properties are formalised and explained in terms of the chosen theory. Typically, there are two kinds of arguments for choosing a specific theory T. Firstly, with theory T we should be able to formalise existing IR models. Secondly, some property P which is shown to be very important for information retrieval purposes should be well-covered by the theory T.

Next, we present the two main types of theories as used in information retrieval, namely those based on probability theory and those based on logic.

Probability Theory One main direction in theoretical information retrieval research is based on probability theory (for an overview of probabilistic IR models see [54]). Typically, in a probabilistic retrieval model one estimates the probability that a user decides that a document is relevant given a particular document and query, denoted as $P(\mathit{Relevant} \mid \mathit{Document}, \mathit{Query})$. Here, an information retrieval theory is centred around the statistical uncertainty assumptions involved in information retrieval. At a meta-level the information retrieval theory could be studied in terms of probability theory.

For instance Cooper [42, 44] inspects some probabilistic assumptions which have consisted of various combinations of the three statistical independence assertions $I1$, $I2$ and $I3$ defined as follows:

$$\begin{array}{lrcl} \text{I1.} & P(A,B) & = & P(A) \times P(B); \\ \text{I2.} & P(A,B \mid R) & = & P(A \mid R) \times P(B \mid R); \\ \text{I3.} & P(A,B \mid \overline{R}) & = & P(A \mid \overline{R}) \times P(B \mid \overline{R}). \end{array}$$

In these formulae A and B are properties of documents or users, depending on the focus of the study. The character R denotes the event of relevance. Assumption I1 reflects the assumption that A and B are independent, which is often assumed to be true in information retrieval for document and information need properties. Assumptions I2 and I3 are adopted by probabilistic model developers [128]. In combination with well-known properties of conditional probabilities, such as $P(A \mid B) = \frac{P(A,B)}{P(B)}$, or phrased differently, $P(A,B) = P(A \mid B) \times P(B)$, assertion $I1$ implies that properties A and B are absolutely independent ($P(A) = P(A \mid B)$ and $P(B) = P(B \mid A)$). Assertion $I2$ and $I3$ express that A and B are independent given relevance or its absence. Stated otherwise, the fact that A is relevant does not influence the fact that B is relevant and vice versa.

On a meta-level Cooper studies the contradiction of elementary laws of probability theories in this information retrieval setting. For instance, using the binary independence retrieval model [54], A is the occurrence of a specific document d and B the occurrence of a specific query q. Then we estimate the probability that document d is judged relevant with respect to query q. Let $P(A) = P(B) = P(R) = 0.1$ and $P(R \mid A) = 0.5$ and $P(R \mid B) = 0.5$. We can calculate that $P(A,B,R) > P(A,B)$[5] which is in conflict with the assumption that a removal of an event always leads to an increase of the probability value. To circumvent this kind of problem, Cooper [44] suggested a reformulation of the underlying assumptions in terms of probability theory. Cooper concludes

> *'When this is done, some models are found to be not only different in character but more realistic than had been supposed, for the true modelling assumptions are weaker and more plausible than the ones thought to be in force.'*

A meta-theory based on probability theory inspects IR models in terms of their uncertainty calculation. The probability calculus is the first-class citizen of this approach. For instance, one can analyse different relevance-functions in terms of a probabilistic inference model as shown in the inference network of Turtle & Croft. Without proposing a new model, Wong & Yao [145] showed that known models such as the boolean, fuzzy set, vector-space, and probabilistic models are special cases of the probabilistic inference models.

Logic The first-class citizens of a logical theory are the inference process and the modelling of information (for an overview of logical IR models see [85]). If a formula φ can be inferred from a formula ψ in a logic L, this could imply that the information represented by ψ is relevant with respect to the information represented by φ. Cooper [41] originated the logical approach by viewing a part of the relevance decision as a logical inference process. Van Rijsbergen suggested that if we are able to infer *relevance* in a logical sense, maybe a particular logic could be used for modelling information retrieval [123, 124]. In a logical theory the study of IR models proceeds by inspecting the logical properties of the retrieval process. One example of using logic in order to analyse information retrieval is given by Chiaramella & Chevallet [35]. They study the semantics of the implication as used in IR models such as the boolean model. In terms of the underlying model they are

[5] According to Cooper, $P(A,B) = P(A) \times P(B) = 0.01$ and $P(A,B,R) = P(A,B \mid R) \times P(R) = P(A \mid R) \times P(B \mid R) \times P(R) = \frac{P(R|A) \times P(R|B) \times P(A) \times P(B)}{P(R)} = 0.025$.

able to propose some extensions based on their logical analysis. The authors conclude that a logical approach 'provides a better way of encompassing the fundamental aspects of information retrieval' [35]. Their conclusion was based on the following three observations. Firstly, the expressive power of the logical model. Secondly, the new insights gained from studying existing models in a logical setting. Lastly, the necessity of coping with the fast change of information retrieval paradigms as presented in Chapter 1.

A theory could also be a combination of two theories, one covering property P very well, the other property Q. For example, in the Logical Uncertainty Principle [124], the combination of a logical and a probabilistic approach is presented. The Logical Uncertainty Principle (which will be explained in more detail in Section 6.2) is founded on the idea that, if an IR system cannot deduce that a document d is about a query q given a logic $\mathcal{L}$, we have to add information to the data set[6] until we can determine the aboutness relation between the document and the query. The strength of aboutness can be associated with the measure of uncertainty $P(d\ about\ q)$ which is based on how much information is added. For example, assume that d is indexed with a logical formula t_1, and that aboutness is defined in terms of classical logic, i.e., if $\vdash d {\rightarrow} q$ then $d\ about\ q$. Then we cannot derive that d is about $t_1 \wedge t_2$. In this particular case we have to add t_2 to d in order to derive aboutness. Applying the Logical Uncertainty Principle one could for instance calculate the uncertainty of t_2 in order to measure $P(d\ about\ t_1 \wedge t_2)$.

A typical example of a model that combines a logical and probabilistic approach is shown in the article 'Towards a Probabilistic Modal Logic for semantic-based Information Retrieval' by Nie [108]. Here he presents an integration of semantic inference (based on a Possible World semantics) and probabilistic measurement based on the Logical Uncertainty Principle.

2.2.4 Theory performance

Various information retrieval experiments have now been around for some time. So far different kinds of experiments have been proposed to determine which retrieval rules are most effective in a general theory.

As we mentioned in Chapter 1, every IR model can be viewed as a theory of relevance, or as Turtle & Croft [143] state it, 'every information system has, either explicitly or implicitly, an associated theory of information access and a set of assumptions that underlie that theory'.

Any proposed IR model is in fact a proposal for a theory of aboutness or relevance between information representatives. The question arises as to how to build a theory for

[6] Van Rijsbergen is not explicit about what we could understand by the concept of a data set. It could be a document d, a query q, or a consulted knowledge-base.

effective information retrieval and, even more importantly, when is one theory preferable over another theory? Hawking [64] defines a good theory as follows:

> *'A theory is good if it satisfies two requirements: it must accurately describe a large class of observations on the basis of a model that contains only a few arbitrary elements, and it must make definite predictions about the results of future observations.'*

Adopting this description we come to the conclusion that a good information retrieval theory should describe relevance decisions in a cognitively acceptable way. It should also present predictions about what happens, for example, if we extend the document-base or change the representations. With the theory in hand, we should also be able to underpin some existing assumptions and hypotheses. For example,

- Information Retrieval is an inference or evidential reasoning process in which we estimate the probability that a user's information need, expressed as one or more queries, is met where a document is taken as 'evidence' (Turtle & Croft [143]).
- We knew that within a single system, it was not possible to improve both the recall and the precision ratio simultaneously, but it was hypothesised that there would be some combination of recall and precision devices which would give optimum performance (Cleverdon [37]).

Our guideline in the comparison of information retrieval theories is the work of the philosopher Popper [117, 118, 119]. In his work Popper proposed a way of developing, and particularly comparing, scientific theories. We believe that in order to analyse, compare, and improve different information retrieval theories a meta-theory is needed.

Popper discussed a general meta-theoretical approach in order to avoid experimental problems. The reason is that the experimental approach eventually leads to the problem of induction, which he calls *Hume's logical problem of induction* [117] and which can be described with the following statements $\mathcal{L}_1, \mathcal{L}_2$ and $\mathcal{L}_3$:

> $\mathcal{L}_1$: Can the claim that an explanatory universal theory is true be justified by 'empirical reasons', that is, by assuming the truth of certain test statements or observation statements (which are 'based on experience')?

In terms of information retrieval, can the experiments mentioned in this chapter, such as the recall and precision measures of a TREC-collection, prove the goodness of a certain IR model?

> $\mathcal{L}_2$: Can the claim that an explanatory universal theory is true or that it is false be justified by 'empirical reasons', that is, can the assumption of the truth of test statements justify either the claim that a universal theory is true or the claim that it is false?

Stated in terms of information retrieval: can experiments prove that a certain IR model is inadequate?

> $\mathcal{L}_3$: Can a preference, with respect to truth or falsity, for some competing universal theories over others ever be justified by such 'empirical reasons'?

Or in terms of information retrieval: can IR models be ordered according to their results of experiments?

With the experimental approach the IR models can indeed be ordered according to their results. However, they are only ordered according to their success and not according to their failure. To clarify this point, consider the following example.

Example 2.3 Assume we want to test the IR models A and B, given a test-collection $\mathcal{D}$. Let the test-query be 'Flying objects without wings'. Let us denote the documents returned by model A and model B by Ret_{A} and Ret_{B} respectively. Then model A has a better recall if and only if $\mid Ret_{\text{A}} \cap Rel_{\text{user}} \mid > \mid Ret_{\text{B}} \cap Rel_{\text{user}} \mid$. Stated differently, the results conforming to the user's relevance decisions are compared. Using Popper's argumentation it is also (and probably even more) interesting to compare $Ret_{\text{A}} \setminus Rel_{\text{user}}$ with $Ret_{\text{B}} \setminus Rel_{\text{user}}$. If, for example, model A assumes 'Planes with wings' relevant and model B does not, the theory behind B could be preferred over the one behind A.

In his work, Popper questioned the validity of proofs by induction for theories. The problem of induction is best formulated in his book 'The myth of the framework' [119], where he gives the following two theses:

- All scientific knowledge is hypothetical or conjectural,
- The growth of scientific knowledge consists of learning from our mistakes.

The failure of an IR model in fulfilling certain requirements will possibly lead towards a better IR model. If we want to build a 'good' or more accurately a 'better' IR model, Popper suggests thirteen steps, which we present in an information retrieval setting[7]. Here, the theory we are searching for is a theory that explains when information in a document d is relevant given a request q.

Step 1 It only makes sense to compare competing theories; that is, information retrieval theories which are offered as solutions to the relevance decision.

Step 2 If we want to create an information retrieval theory we should not only be interested in the truth of 'd is relevant with respect to q', but also in the condition for its falsity because finding that 'd is relevant with respect to q' is false is the same as finding that its negation is true (it is not the case that 'd is relevant with respect

[7] For Popper's original presentation see [117] pages 13–17.

to q'). However we cannot straightforwardly apply a Closed World assumption on an explanatory theory, since the negation of an explanatory theory is not, in its turn, an explanatory theory.

Step 3 We have to search for those cases where a theory breaks down, and not create a new theory that succeeds where its refuted predecessor succeeds, and that also succeeds where its predecessor failed, that is, where it was refuted. If the new theory succeeds in both cases, it will at any rate be more successful and therefore 'better' than the old one. For instance, consider an information retrieval theory that is improved in such a way that it recognises the difference between 'information systems' and 'system information', where the old theory did not, while the rest of the theory remains the same. Now, under the assumption that this recognition is an improvement for information retrieval, we can state that the new theory is 'better' than the old one.

Step 4 If the new theory can handle the problem of the old theory well, and it does not break down in a particular case where the old model broke down, it will be a better explanatory theory.

Step 5 Now we have to search for new cases where the theory can break down, or stated differently, where the decision 'd is relevant with respect to q', fails in the real world but not in the theory.

Step 6 Of course there are several new theories that could handle the break-down case of the old theory, but many of them may be false. The theoretician will therefore try her best to detect any false theory among the set of non-refuted competitors; she will try to 'catch' it. That is, she will, with respect to any given non-refuted theory, try to think of cases or situations in which theory and reality do not agree. Thus she will try to construct severe tests, and critical test situations.

Step 7 By this method of elimination, one may hit upon a true information retrieval theory. However it is not possible to state that this theory is true, that is, that it is the real theory of relevance. The number of possibly true theories remains infinite, at any time and after any number of crucial tests. Maybe among theories actually proposed there is more than one which is not refuted at time t, so that we may not know which of these we ought to prefer. But if at a time t a plurality of theories continues to compete in this way, the theoretician will try to discover how crucial experiments can be designed between them; that is, experiments which could falsify and thus eliminate some of the competing theories. As pointed out in Section 2.1.3, the main goal of TREC can be viewed as the execution of step 7.

Step 8 The procedure described may lead to a set of information retrieval theories. For, although we demand from a new theory that it solves those problems which it predecessor solved and those which it failed to solve, it may of course always happen that two or more new competing theories are proposed such that each of them satisfies these demands and in addition solves more problems than the others.

Step 9 At any time, we are especially interested in finding the best testable of the competing theories in order to submit it to new tests. This will be, at the same time, the one with the greatest information content and the greatest explanatory power. It will be the theory most worthy of being submitted to new tests, in brief it will be 'the best' of the theories competing at time t. If it survives its tests, it will also be the best tested of all the theories considered so far, including all its predecessors.

Step 10 We should take care that our information retrieval theory is not ad hoc, and not create a theory that can only handle particular tests. For instance, consider a test-collection where a large number of the documents is about animals. If we consider in our IR system a descendant system of animals for this specific test we probably obtain a good performance. However this result is dependent on the test-collection.

Step 11 Popper calls this method the *critical method*. It is a method of trial and the elimination of errors, of proposing theories and submitting them to the severest tests we can design.

Step 12 Unfortunately, nothing guarantees that for every theory which has been falsified we can find a 'better' successor, or a better approximation–one that satisfies these demands. There is no assurance that we will be able to make progress towards better theories.

Step 13 The relation between test statements and information retrieval theories may not be as clearcut as is assumed here; or the test statements themselves may be criticised (see Section 2.1.4: this is exactly what happened with the information retrieval test-collections). This is the type of problem which always arises if we wish to apply pure logic to any real world situation. In connection with science it leads to what Popper called *methodological rules*, the rules of critical discussion. The other point is that these rules may be regarded as subject to the general aim of rational discussion, which is to get nearer to the truth.

Ending with Popper's last point, it becomes clear that information retrieval tests and information retrieval theories should be developed in tandem. The continuous search for a 'better' successor of an information retrieval theory is what we call *theory performance*. Theory performance is only possible if one has suitable tests to inspect and perform the theory at one's disposal. Above all, an information retrieval test is only worth considering it if it leads to a better information retrieval theory. The only choice we have to make is which formal tool we are going to use to present the theory. This choice will be made in the next section.

2.3 Situation Theory

As mentioned in Chapter 1, one can divide an information model into two parts, namely the information representation and the matching process. The first part, in its turn, can be divided into document- and query-representations. The formal tool we propose for modelling this kind of information is *Situation Theory* [2, 4, 7, 9, 11, 39, 49]. For the matching process we present a new formal tool based on the representations of information and a logical view of an aboutness proof system in the next chapter.

The reasons for choosing Situation Theory as the ultimate theory for representing information as it occurs in information retrieval are discussed at length in the work of Lalmas [82, 83, 84, 85] and Lalmas & Van Rijsbergen [86, 87, 127]. The approach taken in this thesis differs from theirs in the sense that we view Situation Theory not as a tool to drive information retrieval but as a vehicle to analyse theoretical properties of information retrieval mechanisms. However, we share with them the conviction that Situation Theory presents many characteristics that are both adequate and appropriate for the study of information retrieval. We give a brief overview of the reasons for our conviction based on the article 'Information Retrieval and Situation Theory' [69].

As opposed to classical logic, Situation Theory takes information as the basic, underlying concept, not truth. For instance, a basic activity in classical logic, inference, no longer concerns truth preservation in Situation Theory, but is a form of information extraction and information processing. Situation theory finds its origin in an attempt of Barwise & Perry to create a theory of meaning [10, 11].

In Situation Theory information is represented, not by its *truth* value but by its *content*. A representation of the information content of the document is required, that is, *what* is the information carried by the document, instead of the question whether the information holds in a document (as would be needed in truth predicates in classical-logic-based frameworks). Moreover, if predicates were used to represent the information in the document, many contradictions or intuitively unacceptable deductions could arise, since it can happen that information in a document is by nature logically inconsistent (*'all animals are equal but some animals are more equal than others'*) or in a logical sense meaningless (*'to be or not to be'*). Situation Theory allows us to represent information content. It states that the most important thing is the notion of information, though its precise definition is still a problem.

In information retrieval the content of the complete document can not be represented completely. Due to the fact that an indexing process cannot capture the broad variety of language, some information in the document will not be recorded as a representative of the document. We have to be careful not to assume that if some information is not stored as a representative of a document this will imply that that the negation of the information is inherent in the document. Therefore an information theory should handle partiality in a natural way. Fortunately, Situation Theory does this: if a particular piece of information is not present, then this does not mean in Situation Theory that this information is false. It can be implicit, and some constraints can make one aware of this information.

In most informational frameworks, information is basically represented syntactically. Indeed, a syntax is often proposed that has nothing to do with information content, only with its structure. A semantics is attached to this syntax so that it can model the information content. In Situation Theory, however, the semantics is explicitly incorporated as a first-class citizen. There is no distinction between syntax and semantics. A syntax is used so semantics can be expressed.

The use of Situation Theory to develop a meta-theory for information retrieval leads us to a better understanding of the nature of information in information retrieval. Choosing this theory we can also look at the nature of information for user modelling. A correct representation of the user's intention certainly generates better retrieval. The attainment of such a representation enters the area of Cognitive Science, some aspects of which can be formally expressed with Situation Theory.

2.4 Summary and conclusions

In this chapter we have presented two different possibilities for studying information retrieval, namely an experimental one and a theoretical one. In the first approach an answer to an information retrieval research topic is validated by means of experiments. In the latter approach the solution has to be proved based on a certain theory. In this thesis the theoretical approach for studying information retrieval is chosen. Furthermore we presented the ideas of Popper. Through his thirteen steps, Popper showed how the validity or failure of proofs can possibly lead towards a better IR model. Finally, we briefly present the reasons for choosing Situation Theory to be the underlying theory of information for our framework.

Chapter 3

The framework

What do we then but draw anew the model
In fewer offices, or at last desist
To build at all? Much more, in this great work,
Which is almost to pluck a kingdom down
And set another up, should we survey
The plot of situation and the model,
Consent upon a sure foundation,
Question surveyors, know our own estate.

W. Shakespeare, *'Henry IV – part 2'*.

In the introduction of this thesis, we stated that information retrieval concerns the problem of retrieving from a given document-base those documents that are *likely* to be *relevant* to a certain information need. In 1971, Cooper introduced an objective part of the relevance relation termed *logical relevance* [41]. We call this relation *aboutness*.

The previous chapter introduced the reasons for developing a general framework for studying the *aboutness* relation. In this chapter we propose such a framework, which captures all concepts necessary to study aboutness as used in information retrieval.

Although there is no consensus about paradigms, and on what is considered information retrieval and what is not, there seems to be general agreement that an IR model can be decomposed into three components, namely:

- a model for the documents;
- a model for the queries;
- an aboutness relation between the two component models.

With this decomposition in mind, we develop a theoretical framework that can be used to study each of the three components. Within this framework it is possible to examine different aboutness relations and study them inductively. Although it is possible to investigate the aboutness relation in isolation, it is also important to consider the underlying document and query models. We start with document models and analyse them from a situation-theoretic perspective. As it turns out, this framework offers the freedom to explore issues relating to document models in a neutral setting.

If we want to study the aboutness relation in a meta-theory, we have to be clear on the domain of the aboutness relation. Is it a relation between models and formulae? Is the relation an association between a set of sentences and a sentence? What we need is an information retrieval-theoretic study of aboutness. Probably, we will never be able to completely formalise the aboutness decisions that humans are capable of, but the study of aboutness can systematise our implicit understanding of this human *aboutness* behaviour and clarify some of the underlying assumptions.

In our view, the aboutness relation is an association between types of information. In this chapter we present a framework that allows us to study aboutness as such an association (see [26, 67]). In Section 3.1 we present the basic concepts of *Situation Theory*. Section 3.2 introduces the information retrieval representatives in terms of an underlying framework. In Section 3.3 we formulate postulates that describe some properties of aboutness as used in information retrieval. In addition to the aboutness postulates, Section 3.3.2 presents some anti-aboutness postulates, which describe some properties of the opposite of aboutness. In order to combine different IR models, some postulates are proposed that describe several combinations. Section 3.4 concludes this chapter with a brief summary.

3.1 Situation Theory

As mentioned in Chapter 2, our view on information is based on Situation Theory. The situation theoretical approach starts with the work of Dretske [50] who presents a philosophical view of information. In his words, there is a signal between the sender and the receiver. The signal may have a meaning, that is, what the sender intended by sending it. More importantly a signal always carries information, or as Dretske puts it: 'What information a signal carries is what it is capable of *telling* us truly, about another state of affairs.' The extraction of the information carried by a signal is viewed as a digitalisation process, i.e., *'a conversion of information from analog to digital form'*.

Analog information is considered to be information carried by the signal. An unspecified agent perceives the signal, by way of some sensor, seeing, feeling, smelling, hearing, etc. This stage is referred to as *perception*. The next stage, *cognition*, involves the extraction of specific items of information from this perceived 'continuum', i.e., the conversion from analog to digital information.

Situation Theory, introduced in the early nineteen eighties, is a mathematical theory of information based on Dretske's view of information [2, 4, 7, 9, 11, 39, 49]. In Dretske's terminology *'a signal can carry the information that s is F'* (where s denotes some item at the source and F an item of information). In Situation Theory *'a signal carries the information that a situation s supports the item of information F, or stated otherwise, that situation s is of some type indicated by F'.*

The primitives of Situation Theory are *situations* which stand for events, properties and relations. Devlin [49] puts this as follows:

> 'the behaviour of people varies systematically according to the kind of situation they are faced with: threatening situations, spooky situations, pleasant situations, challenging situations, conversation situations, and what-have-you, all evoke quite different responses.'

In the theory, situations are partial descriptors of the real world. Situations can also be elements of situations, standing in relation to each other and to other things.

The types of situations, originally named *states of affairs* and by Devlin [48] introduced as *infons*, take the form of collections of basic facts. Infons are considered as properties holding for situations. Information is not represented by the truth value of the infons but by the truth value of the proposition 'infon φ holds in situation S'. The notion of *holds in*, often referred to as *the support relation*, is denoted as $\models$. For instance, given an infon φ and a situation S, the proposition $S \models \varphi$ means that the information item φ holds in situation S, or stated differently, that situation S supports infon φ.

A more formal definition of an infon is needed to work with. Devlin [49] defined the notion of an *infon* as follows:

Definition 3.1 An infon is an item $\langle\langle R,\mathrm{a}_1,\ldots,\mathrm{a}_n;\ i\rangle\rangle$ that represents that the relation R holds (if $i = 1$) or does not hold (if $i = 0$) between the objects $\mathrm{a}_1,\ldots,\mathrm{a}_n$.

The objects in this definition[1] include the following: *individuals*, such as 'John', 'table', etc.; *spatial locations*, such as, 'garden','here', etc.; *temporal locations*, such as '10am','now', etc.; *situations*, some structured parts of the world as discussed before; *types*, high order uniformities, for instance the situation types (see later); and *parameters*, indeterminates in the definition that range over objects of the various types, denoted by $\dot{p}$,$\dot{q}$, $\dot{p}_1,\ldots,\dot{p}_n$.

The relation R is a uniform property that holds of, or links, the objects. The value i is called the *polarity* of the infon. If the polarity is 1, we call the infon *positive*; it is called *negative* otherwise.

[1] Devlin's article 'Infons and Types in an Information-Based Logic' [48] presents the definition of infons and types in all detail.

Let us illustrate these notions with an example. Consider the situation S presented in the book entitled *Julius Caesar* by Shakespeare, in which Caesar dies. Any person reading this part of the book is able to extract information from it, such as: 'Who killed Caesar?', 'Where is Caesar killed?', etc. For instance if the reader understands that 'Brutus killed Caesar', this can be modelled by the infon $\langle\langle \textit{Killed}, \text{Brutus}, \text{Caesar}; 1\rangle\rangle$. The relation *Killed* holds between the individuals "Brutus" and "Caesar". The proposition $S \models \langle\langle \textit{Killed}, \text{Brutus}, \text{Caesar}; 1\rangle\rangle$ is true, it provides us with information about the document. Reading this part of the document, one may argue that the situation also supports the negative infon $\langle\langle \textit{Killed}, \text{Caesar}, \text{Brutus}; 0\rangle\rangle$. It does not support the infon $\langle\langle \textit{Killed}, \text{Cain}, \text{Abel}\ ;1\rangle\rangle$, since this infon certainly cannot be perceived from the situation S. Whether this infon holds or does not hold in general is of no importance: it does not hold in situation S.

If one of the objects used in an infon is a parameter, it is called a *parametric infon*. A parameter in an infon is used to express that a reference should be linked to an arbitrary object. In order to extract information from the proposition $S \models \langle\langle$parametric infon$\rangle\rangle$, the parameters of the infon have to be instantiated. This process is referred to as *anchoring*. For instance, the proposition $S \models \langle\langle \textit{Killed}, \text{Brutus}, \dot{\text{p}}; 1\rangle\rangle$ does not provide us with information about S unless there is an anchor from $\dot{\text{p}}$ to an individual (for example "Caesar" or "Cleopatra"). So, only when $\dot{\text{p}}$ is anchored to some specific person does the proposition $S \models \langle\langle \textit{Killed}, \text{Brutus}, \dot{\text{p}}; 1\rangle\rangle$ provide us with information about the document. For this reason Devlin views a parametric infon as a kind of 'template' for an item of information.

Obviously it is hard to formally decide whether some situation supports a given infon. Above all it is difficult to represent all the infons that are supported by a situation. In order to create a mathematical theory, Situation Theory distinguishes two types of situations, namely *real situations* and *abstract situations*. A real situation is referring to the real world, an abstract situation is a mathematical construct, consisting of a set of infons.

The support relation for an abstract situation can easily be formulated, based on set-theoretic membership:

Definition 3.2 The binary relation *supports* between an abstract situation S and an infon φ, denoted as $\models$, is defined by:

$$S \models \varphi \Leftrightarrow \varphi \in S$$

Types in Situation Theory are 'higher order uniformities'[2]. Consider the two infons $\langle\langle \textit{Killed}, \text{Brutus}, \text{Caesar}; 1\rangle\rangle$ and $\langle\langle \textit{Killed}, \text{Brutus}, \text{Cleopatra}; 1\rangle\rangle$. These infons are essentially

[2] For detailed information see [49], page 50.

describing a situation in which 'Brutus killed someone'. The only difference is whom is killed by Brutus. For the two infons there is a unifying type, namely:

$$\tau = [S \mid S \models \langle\langle \mathit{Killed}, \text{Brutus}, \dot{\text{p}};\ 1\rangle\rangle]$$

The higher-order uniformity presented, is that a situation S is of type τ if and only if in this situation 'Brutus killed someone', that is, if and only if $S \models \langle\langle \mathit{Killed}, \text{Brutus}, \dot{\text{p}};\ 1\rangle\rangle$ and $\dot{\text{p}}$ can be anchored to some specific person. The type τ is an example of a so-called situation-type. In Situation Theory other so-called basic types are considered, for example, the type of a temporal location, the type of a spatial location, etc. We refer the reader to [48] for a complete and in-depth presentation of types as used in Situation Theory.

In the following section we model the information retrieval concept of information in terms of Situation Theory. An attentive reader might wonder to what extent the rest of this thesis depends on our choice of Situation Theory. One might for instance suggest to use Possible World Semantics as the basis for a theory of information, rather than the apparently more esoteric Situation Theory. We would like to emphasise that the axiomatic (or logical) approach to aboutness which we present in Section 3.3 does not depend on the choice of the representation of information. Furthermore, the choice of another representation of information does not have to exclude our framework. We are aware of the fact that several authors have inspected the relation between Situation Theory and other (information) theories [12, 56, 133, 147]. For instance, Zalta [147] answered the question whether Situation Theory and World Theory (such as Possible World semantics) could peaceably coexist. Zalta proposed an assimilation of infons, situations and worlds into a single axiomatic theory that distinguishes and comprehends all three kinds of entity. In his theory twenty-five theorems are proposed which are 'basic, reasonable principles that structure the domains of properties, relations, states of affairs [infons], situations, and worlds in true and philosophically interesting ways'. Still, we agree with the arguments given by the developers of Situation Theory, namely that an information theory should model 'information' rather than 'truth'.

3.2 Modelling information retrieval concepts using Situation Theory

In our framework, the sender of the signal that carries information could be viewed as an author who wants to inform the reader (as the receiver) in some way or another. Here, the signal that carries information is termed a *document*, be it a book, a movie, pictures, etc. In documents, various situations are present. In order to have a running example at our disposal we present a small document.

Example 3.1 Consider the following document: 'The Sioux defeated General Custer's cavalry at Little Big Horn in 1876. Nobody knows what really happened in the battle. Some historians believe that the Sioux heavily outnumbered Custer's men.'

Reading this document, we are able to individuate several items of information. More specific, every reader can individuate (perceive) different items of information. Some might say that this document is best represented by the information item "battle", others would say that the best representative information item should be "Sioux".

We propose that in the scope of information retrieval the perception of a reader or agent takes place by a special variant of the sensor seeing, namely *reading*. The digitalisation of information takes place by the representation of documents into representative information items. In practice, the agent is a computerised indexing system.

The process of indexing involves a huge loss of information. One of the reasons is that the digitalisation of information takes place in terms of a representation language that cannot capture the broad variety of information for the reasons mentioned in Chapter 1.

So far, no mention has been made of queries. A query is a request for information, by means of which the user supplies the information items that supposedly represent the information she is interested in sufficiently closely. As a query can thus be seen as a set of information items, we do not distinguish the information corresponding to an information need and the information in a document.

Using a representation language one can represent representative information items inherent in a document as a set of descriptors or, in Situation Theory, infons. A representation of the document consists of a set of descriptors. These descriptors can be almost anything, for instance, keywords, boolean formulae, conceptual graphs, photo's, noun phrases, etc. One can see the representation of a book as an abstract situation, constructed from the real situations presented in the book.

Since we attempt to define a formal framework for information retrieval, we focus on the mathematical representation of situations, or stated differently, on abstract situations. What is needed are infons suitable to model the information retrieval descriptors of different representation languages. First we introduce a specific kind of infons, the *profons*.

3.2.1 Profons

Keywords play a pivotal role in the majority of representation languages. In efficient algorithms which automatically index keywords from the document one often uses representation languages that contain only keywords descriptors. As a result, the relationships in which the keywords stood are not included. Restoring and detecting these relationships automatically and efficiently is an arduous task. The ability to deal with (and thus to index) large document collections is seen by many researchers as the ultimate goal for information retrieval. The majority of the IR models, for instance, the probabilistic and vector-space ones, use the keyword characterisation of information for their aboutness decision.

Using keywords to model information results in very simple infons. In a sense, keyword-based infons can be considered 'sub-informational' particles, just as protons are to atoms. For this reason we introduce the term *profon* [67]. Profons are infons based on an unspecified unary relation. This relation (denoted with I) reflects the fact that all knowledge of the relations that the keyword was part of is abandoned. If a document d is represented by the keyword "Sioux", one may conclude that the information conveyed by "Sioux" is inherent, or holds in, document d.

In our framework, profons are intended to capture the *basic information items* present in a document. These basic information items can be much more than just keywords, for instance they can be noun phrases. Thus, the item "Little Big Horn" is considered to be a basic information item and $\langle\langle$I,Little Big Horn; 1$\rangle\rangle$ is the corresponding profon. Throughout this thesis the set $\mathcal{T}$ is used to denote a finite set of basic information items $\{t_1, \ldots, t_n\}$. By putting basic information items into Situation Theory terminology, we can now formally introduce the notion of profon.

Definition 3.3 The language $\mathcal{P}(\mathcal{T})$ of profons is defined by:

$$\mathcal{P}(\mathcal{T}) \ =^{def} \{\langle\langle \text{I,t; j}\rangle\rangle \mid t \in \mathcal{T}, j \in \{0,1\}\}.$$

Typical elements of $\mathcal{P}(\mathcal{T})$ will be denoted as $p, p_1, \ldots, p_n$. Representation languages that contain only positive profons, i.e., profons with polarity 1, can be viewed as subsets of $\mathcal{P}(\mathcal{T})$. The full subset of positive profons is denoted by $\mathcal{P}^+(\mathcal{T})$ and consists of all elements $\langle\langle$I,t; 1$\rangle\rangle$. Whenever the set $\mathcal{T}$ is understood we write $\mathcal{P}$ rather than $\mathcal{P}(\mathcal{T})$. For the sake of brevity, we denote positive profons without mentioning the relation and polarity. For example, the profon $\langle\langle$I,Sioux; 1$\rangle\rangle$ is denoted by $\langle\langle$Sioux$\rangle\rangle$.

3.2.2 Infons

A feature of information items is that they can be manipulated to form more *complex* information items. For example, two pieces of information can be combined to form a new piece of information. The combination of two infons should result in an infon. So far, we have only introduced profons. In order to model combined information items, more complex infons are needed. The language $\mathcal{I}(\mathcal{P}, Rel, Prm)$ which is the language of infons, is defined as follows:

Definition 3.4 Let $\mathcal{P}$ be the profon language as in Definition 3.3, Rel a finite set of relations and Prm a set of parameters. The language $\mathcal{I}(\mathcal{P}, Rel, Prm)$ of infons is defined to be the smallest superset of $\mathcal{P}$ such that

- If $a_1, \ldots, a_n \in \mathcal{I}(\mathcal{P}, Rel, Prm) \cup Prm$ and $R \in Rel$ is a n-place relation then $\langle\langle R, a_1, \ldots, a_n;\ 1\rangle\rangle \in \mathcal{I}(\mathcal{P}, Rel, Prm)$ and $\langle\langle R, a_1, \ldots, a_n;\ 0\rangle\rangle \in \mathcal{I}(\mathcal{P}, Rel, Prm)$

for any $n \in \mathbb{N}$ with $n > 0$.

Typical elements of $\mathcal{I}(\mathcal{P}, Rel, Prm)$ will be denoted as $\varphi, \psi, \psi_1, \ldots, \psi_n$. Whenever the sets $\mathcal{T}$, Rel and Prm are understood we write $\mathcal{I}$ rather than $\mathcal{I}(\mathcal{P}, Rel, Prm)$. Devlin [49] speaks of infons constructed from a given set of basic information items as *compound infons*, as their existence is due to an information combination.

3.2.3 Relations

In information retrieval some indexing processes combine information items by bringing them into a relationship, in order to describe the information content of a document more precisely. Take, for example, the keywords "Custer" and "Adventures". These can be combined to form the phrase "Custer's adventures" or "Adventures of Custer". In the framework we can model such a compound information item through the infon $\langle\langle Possession, \langle\langle \text{Custer}\rangle\rangle, \langle\langle \text{Adventures}\rangle\rangle;\ 1\rangle\rangle$. This infon is the result of combining the two profons $\langle\langle \text{Custer}\rangle\rangle$ and $\langle\langle \text{Adventures}\rangle\rangle$ indicating that the profons are associated by the relation *Possession*, drawn from a predefined set of relations. This set of predefined relations can vary from language to language. The most common example is the set of logical relations used in the boolean model.

Boolean relations

In the boolean model we have the logical connectives $\wedge$, $\vee$ and $\neg$. Given these connectives we can define the boolean infon language as follows:

Definition 3.5 The *boolean infon language* $\mathcal{I}_{Bl}(\mathcal{T})$ is the infon language $\mathcal{I}(\mathcal{P}(\mathcal{T}), \{\wedge, \vee, \neg\}, \emptyset)$.

For example, the infon $\langle\langle \wedge, \langle\langle \text{Sioux}\rangle\rangle, \langle\langle \text{Cavalry}\rangle\rangle, \langle\langle \text{Sleep}\rangle\rangle;\ 1\rangle\rangle$ expresses the informational composition of the profons $\langle\langle \text{Sioux}\rangle\rangle$, $\langle\langle \text{Cavalry}\rangle\rangle$ and $\langle\langle \text{Sleep}\rangle\rangle$. Intuitively, this compound infon describes the existence of information about the three given information items in a situation, but this does not need to be directly related information. For example, "The Sioux were hunting for food. The cavalry was sleeping" is a valid situation supporting the compound infon. A more sophisticated indexing approach that conserves the informational relatedness between information items can be found in Farradane's relational indexing.

Farradane's relational indexing

In Farradane's work [51, 52] information is carried by a fixed set of relationship types over an underlying set of terms. It is based on the idea that much of the meaning of information objects is encapsulated in the relationships between terms. Farradane proposed a set of nine primitive relationship types through which any given term relationship could be classified (see Figure 3.1). He motivated his relationship types on the basis of psychological thought mechanisms. In his relational indexing, trained indexers (humans!) would take an object and classify the term relationships[3].

		Associative mechanisms		
	Conceptualisation	Awareness	Temporary association	Fixed association
	Concurrent concept	**Concurrence**	**Self-activity**	**Association**
Discriminatory mechanisms	Not-distinct concept	**Equivalence**	**Dimensional**	**Appurtenance**
	Distinct concept	**Distinctness**	**Action**	**Functional dependence**

Figure 3.1: Farradane's nine relationship types.

We can construct a language of infons $\mathcal{I}_{Far}(\mathcal{T})$ in which the set of relations is given by the nine relationship types of Farradane.

Definition 3.6 The *Farradane relation infon language* $\mathcal{I}_{Far}(\mathcal{T})$ is the infon language $\mathcal{I}(\mathcal{P}^{+}(\mathcal{T}), \{\text{Concurrence, Equivalence, } \ldots \text{, Functional dependence}\}, \emptyset)$.

For example, the infon $\langle\langle \mathit{Concurrence}, \langle\langle \text{Custer} \rangle\rangle, \langle\langle \text{Cavalry} \rangle\rangle; 1 \rangle\rangle$ expresses that information about "Custer" appears in the presence of information about "cavalry" (expressed linguistically also as "Custer's cavalry"). Even though the infons in this language clearly capture more of the content of an object than the profons presented so far, the disadvantage is that indexing has to be performed manually and is not driven by a formal specification, which makes it hard to decide for example that a term is in *Concurrence*-relation with another term.

Index expressions

In Bruza's work [23], a practical variant of Farradane's approach is proposed, the so-called index expressions. An index expression consists of a number of terms, separated by means

[3] For Farradane's detailed presentation and argumentation for choosing these relations see [51, 52].

of connectors modelling the relationships between these terms. Terms are taken from a given set $\mathcal{T}$ of terms and correspond to nouns, noun-qualifying adjectives and noun phrases; connectors are taken from a set C of connectors and are basically restricted to prepositions in addition with a so-called null connector $\circ$ to express term-phrases such as Little Big Horn. For example, the proposed connector set of Bruza contains the elements, $\{\circ$, "about","and", ..., "with","within","without"$\}$ (for detailed information see [23]).

The advantage of this approach is that the indexing process can be performed automatically (for details see [23]). The disadvantage is that one is not able to express the similarity between the situations of the type "the murder of Caesar" and "Caesar's murder".

Definition 3.7 The *index infon language* $\mathcal{I}_{Idx}(\mathcal{T})$ is the infon language $\mathcal{I}(\mathcal{P}^+(\mathcal{T})$, $\{\circ$,of,in,at,...,around$\}, \emptyset)$.

Given this index infon language, the infon $\langle\langle in, \langle\langle \text{Battle}\rangle\rangle, \langle\langle 1876\rangle\rangle; 1\rangle\rangle$ expresses information about a "Battle in 1876".

So far we presented document descriptors as profons and infons. Next, we focus on two other aspects relating to document descriptors, namely *information containment* and *preclusion*.

3.2.4 Information containment

In information retrieval it can be of use to infer additional information, information that is implicit in the infon that is given, in order to use it for the aboutness decision. We have some underlying assumptions for the inference of information which depend on the IR system under consideration. For instance, in the boolean system, we have that from the proposition $p \wedge q$ it is possible to infer p or even $p \wedge q \wedge \neg r$ if the Closed World assumption is adopted. In this case the information inference is based on a logical deduction system. In other models other inference rules are used. In the situated information retrieval framework we also have the notion of information inference.

Commonly, in information retrieval, information inference is based on the notion of information containment [26]. For instance, from the infon $\langle\langle \wedge, \langle\langle \text{Sioux}\rangle\rangle, \langle\langle \text{Cavalry}\rangle\rangle; 1\rangle\rangle$ the profon $\langle\langle \text{Sioux}\rangle\rangle$ can be inferred, as the latter infon is informationally contained in the former. According to Barwise & Etchemendy [8], infons can be partially ordered with respect to information containment (denoted by $\rightarrow$). In information theory it is assumed that the relation $\rightarrow$ is reflexive ($\varphi \rightarrow \varphi$), anti-symmetric (if $\varphi \neq \psi$, then at least one of $\varphi \not\rightarrow \psi$ or $\psi \not\rightarrow \varphi$), and transitive (if $\varphi \rightarrow \psi$ and $\psi \rightarrow \gamma$ then $\varphi \rightarrow \gamma$). This last property is also referred to as the Xerox Principle, which originates from Dretske [50]).

The ordering with respect to information containment in our framework depends on the information containment relation of the underlying IR model. For instance, in boolean models $\psi \rightarrow \varphi$ is defined by $\psi \vdash \varphi$.

The properties of the → relation for the infons in the boolean infon language $\mathcal{I}_{Bl}(\mathcal{T})$ could be defined straightforwardly, for any ψ_j an infon and $1 \leq j \leq n$ and $j \leq k$ as follows:

$$\langle\langle\wedge,\psi_1,\ldots,\psi_j,\ldots,\psi_n;\ 1\rangle\rangle\rightarrow\psi_j\rightarrow\langle\langle\vee,\psi_1,\ldots,\psi_j,\ldots,\psi_k;\ 1\rangle\rangle.$$

For relations in other languages such as the Farradane infon language $\mathcal{I}_{Far}(\mathcal{T})$, things are not that clear. Often $\langle\langle R,\psi_1,\ldots,\psi_j,\ldots,\psi_n;\ 1\rangle\rangle\rightarrow\psi_j$ holds for any $1 \leq j \leq n$, but it depends on the definition of the underlying relation R.

3.2.5 Preclusion

Of course, not all infons can be meaningfully combined. The reason for this is that the information present in the infons can be contradictory. In this case, infons φ and ψ are said to preclude each other, denoted by $\varphi\perp\psi$. It is natural to assume that an infon with polarity 1 precludes the same infon with polarity 0. The notion of information preclusion is considered fundamental to a theory of information [89]. Seligman [135] considers the preclusion relation as a 'negative' constraint between information items[4], in contrast with the positive constraint →.

Preclusion is interesting for information retrieval because if it is known that two infons preclude each other, then this may be used to determine aboutness[5] [26]. For example, if a document is characterised with the infon $\langle\langle$*defeated*,$\langle\langle$Sioux$\rangle\rangle$,$\langle\langle$Cavalry$\rangle\rangle$; 1$\rangle\rangle$ and the assumption is made that this infon precludes the infon $\langle\langle$*defeated*,$\langle\langle$Cavalry$\rangle\rangle$,$\langle\langle$Sioux$\rangle\rangle$; 1$\rangle\rangle$, then we may be able to derive (for instance by default) that the document also contains the information that $\langle\langle$*defeated*,$\langle\langle$Cavalry$\rangle\rangle$,$\langle\langle$Sioux$\rangle\rangle$; 0$\rangle\rangle$. Therefore this document can be relevant for somebody who is looking for information about the fact that *'The cavalry did not defeat the Sioux'*.

3.2.6 Situations

Infons constitute the lowest level of information granularity. At a higher level of granularity we find the abstract situations, or in information retrieval terminology, the document representation and queries [67, 71].

We can combine abstract situations, which actually are sets of infons, with the normal set operators ∩ and ∪. For example, an encyclopaedia can be viewed as a complicated description of a diverse range of situations and is represented by the situation S_E. This situation S_E consists of a union of unrelated situations. One situation S_F may supports the infon $\langle\langle$France$\rangle\rangle$, which can be seen as the description given for the word "France".

[4] Actually, Seligman considers the preclusion as a relation between types, rather than infons.

[5] Or more precisely, non-aboutness.

Another situation S_S supports the infon $\langle\langle \text{Sioux} \rangle\rangle$, the description of the “Sioux”. The situation S_E is a union of S_F and S_S and all the other situations supporting the other items of the encyclopedia.

However, sometimes it is necessary to put situations in relation with each other. By way of illustration, consider two situations, one in which “Custer’s cavalry was defeated” and another in which “the Sioux defeated someone”. It is possible to create new information using these situations by stating that “the Sioux defeated Custer’s cavalry”. This is based on the assumption that “Custer’s cavalry” is the someone in the situation “the Sioux defeated someone”. This is an example of *situation fusion*, as the respective situations are composed very tightly.

There are different ways to define this kind of situation fusion. Situation fusion is modelled by an operator which, given two situations S and T, results in a situation U (denoted by $S \odot T = U$). One way of defining *situation fusion* is to compose each infon of the first situation with all the infons of the second situation. Another definition can be obtained by the composition of particular infons. For example, take the situation $S = \{\langle\langle \textit{defeated}, \dot{p}, \dot{q};\ 1 \rangle\rangle, \langle\langle \textit{group}, \langle\langle \text{Sioux} \rangle\rangle, \dot{p};\ 1 \rangle\rangle\}$ and the situation $T = \{\ \langle\langle \textit{defeated}, \dot{r}, \dot{s};\ 1 \rangle\rangle$, $\langle\langle \textit{group}, \langle\langle \text{Cavalry} \rangle\rangle, \dot{s};\ 1 \rangle\rangle\}$. If we want to make clear that “the Sioux defeated the cavalry”, we have to state that in the union of S and T the parameters $\dot{p}$ and $\dot{r}$ (respectively $\dot{q}$ and $\dot{s}$) are the same. Note that the same result can be achieved using a union-operator and a correct choice of the parameters in both sets. The fusion process is based on semantical information concerning the two infons of the two situations. Therefore this can hardly be defined in general. In this thesis we will not use the notion of fusion since we will be able to represent all situation combinations of the IR models presented in this thesis using the union and intersection-operators.

A final aspect of situations is the question when two situations can be considered to be identical. As an abstract situation is represented as a set of infons, the meaning of the situation does not depend on the order of the infons and thus situation equality is essentially an instance of set equivalence. Set equivalence is denoted by $\equiv$. For situations S and T the statement $S \equiv T$ intuitively means that the information of S is equal to the information of T and vice versa. In case parameters are used we could say that two situations are equivalent if and only if the two situations can be made textually equivalent by renaming the parameters. Such a renaming is used in lambda calculus and is called α-conversion. Expressions that can be made textually equivalent are called α-convertible.

Our formalisation of information as used in information retrieval started with the profons to model basic information items and ended with situations to model the complete document contents. Formally, the language $\mathcal{S}(\mathcal{I})$ which is the language of situations, is defined as follows:

Definition 3.8 Let $\mathcal{I}$ be the infon language as in Definition 3.4. The language $\mathcal{S}(\mathcal{I})$ of situations is defined to be the powerset of $\mathcal{I}$.

Typical elements of $\mathcal{S}(\mathcal{I})$ will be denoted as $S, T, U, V, S_1, \ldots, S_n$. Whenever the language $\mathcal{I}$ is understood we write $\mathcal{S}$ rather than $\mathcal{S}(\mathcal{I})$.

3.3 The aboutness proof system

In order to create a platform for a discussion about aboutness decisions, we have to make some explicit assumptions of aboutness. The first of these is that aboutness can be derived with some sort of *logic*.

Formally, we represent the aboutness relation between situations S and T with the symbol $S \mathrel{\Box\!\!\rightsquigarrow} T$: intuitively $S \mathrel{\Box\!\!\rightsquigarrow} T$ means that situation S is about situation T, and $S \mathrel{\Box\!\!\not\rightsquigarrow} T$ that situation S is not about situation T. In conformity with reality it is often not immediately clear whether a situation S is about another situation T. We suggest that aboutness can be more or less logically derived [71]. These logical derivations play an important role both in information retrieval as well as in Situation Theory [8].

For instance, given the fact that $S \cup T$ is about S, we can derive that situation $\{\langle\langle$*defeated*,$\langle\langle$Sioux$\rangle\rangle$,$\langle\langle$Cavalry$\rangle\rangle$; 1$\rangle\rangle$, $\langle\langle$*fought*,$\langle\langle$Sioux$\rangle\rangle$,$\langle\langle$Cavalry$\rangle\rangle$; 1$\rangle\rangle\}$ is about the situation $\{\langle\langle$*defeated*,$\langle\langle$Sioux$\rangle\rangle$,$\langle\langle$Cavalry$\rangle\rangle$; 1$\rangle\rangle\}$. This kind of derivations are used to model the aboutness decision of IR models. In most cases an aboutness relation can be described by an effectively given set of axioms and rules. This is for instance the case for derivation in classical proposition logic, but also for derivation in several modal logics [65]. First we define the language needed for the aboutness proof system.

Definition 3.9 For a given infon language $\mathcal{I}$, the *aboutness language* $\mathcal{L}(\mathcal{I})$ of aboutness formulae is the smallest set such that

- if $\varphi, \psi \in \mathcal{I}$ then $\varphi \rightarrow \psi, \varphi \perp \psi, \psi \not\rightarrow \varphi,\ \varphi \not\perp \psi \in \mathcal{L}(\mathcal{I})$;
- if $S, T \in \mathcal{S}(\mathcal{I})$ then $S \mathrel{\Box\!\!\rightsquigarrow} T, S \mathrel{\Box\!\!\not\rightsquigarrow} T, S \equiv T, S \not\equiv T \in \mathcal{L}(\mathcal{I})$

where $\mathcal{S}(\mathcal{I})$ is as in Definition 3.8.

Typical elements of $\mathcal{L}(\mathcal{I})$ will be denoted as $\Psi, \Phi, \Phi_1, \ldots, \Phi_k$. Whenever the language $\mathcal{I}$ is understood we write $\mathcal{L}$ rather than $\mathcal{L}(\mathcal{I})$.

The definition of the aboutness proof system is the following.

Definition 3.10 An *aboutness proof system* is a triple $\mathcal{A}_{ps} = \langle \mathcal{L}, Ax, Rule \rangle$, where:

- $\mathcal{L}$ is an aboutness language as in Definition 3.9;
- Ax is a decidable subset of $\mathcal{L}$, the elements of which are called axioms;

- $\mathit{Rule} = \{R_1, \ldots, R_k\}$ is a finite set of rules of the form $R(T_1, \ldots, T_k, T_{k+1})$ with for each $1 \leq i \leq k+1$, $T_i \in \mathcal{L}$. Here, $T_1, \ldots, T_k$ are the premises of the rule and T_{k+1} is the conclusion. We assume that each R_i is decidable as a relation.

Note that we do not make any statement about how the derivation relation is being determined by the proof system. For example, it is possible that this is analogous to the classical logical derivation: an aboutness decision is derivable from another one if there is a range of 'intermediate' decisions, which are either an axiom or arise from previous decisions by application of a rule. However, one can also think of another proof system. It is for instance possible to consider a default theory (cf. [121]) as a proof system. In this case the derivation relation is defined by being an element of an extension; this derivation relation is in general not expressible in the way of the classical logical derivation relation. Generally, this is the case for non-monotonic derivation relations [94].

In this thesis we assume that a proof of aboutness is a finite-length sequence of aboutness formulae that are either axioms of the derivation system or conclusions of rules applied to formulae that appear earlier in the sequence.

This aboutness proof system results in a sufficiently abstract framework in which the inference mechanism of an arbitrary retrieval mechanism can be captured, and maps it to inference between aboutness relation of situations.

If we build an aboutness proof system $\mathcal{A}_{ps}$ out of aboutness axioms and rules, theorems will be aboutness assertions in the language $\mathcal{L}$. These theorems are all elements of the language $\mathcal{L}$, which are provable in a given deduction system $\mathcal{A}_{ps}$. Alternatively we express that a theorem Φ is provable in system $\mathcal{A}_{ps}$ as $\vdash_{\mathcal{A}_{ps}} \Phi$. As mentioned, we are especially interested in theorems of the form $S \,\Box\!\!\rightsquigarrow T$, which we call *Aboutness Theorems.*

3.3.1 Reasoning with situation aboutness

In our theory *aboutness* is treated as a relation between situations. Therefore aboutness is treated as a fundamental notion with regard to information. This differs from other approaches [23, 84], in which aboutness can be expressed in terms of so-called information containment. In this section a set of postulates is presented consisting of a series of axioms and rules which establishes properties of the aboutness relation between situations. Note that the axioms should not be interpreted as an absolute truth in all cases. We will see later that some axioms are not universally valid but only hold within the context of a particular retrieval system. This offers the possibility to compare retrieval systems according to which axioms and rules they satisfy.

The intuitive interpretation of the rules is as usual in logical systems, i.e., $\dfrac{A}{B}$ means that if A is valid in an IR model, then B is also valid.

Basic postulates

The first axiom, Reflexivity, expresses that any situation is about itself. Reflexivity seems to be an inherent property of aboutness in many IR models.

Reflexivity (Re)

$$S \,\Box\!\!\rightsquigarrow S$$

Note that with this axiom we have that $\emptyset \,\Box\!\!\rightsquigarrow \emptyset$. Sometimes this is an undesirable property, for instance, if one wants to exclude aboutness decisions involving an empty-set. In order to avoid this kind of deductions one could adopt a special version of the Reflexivity axiom called Singleton Reflexivity, which is defined as follows:

Singleton Reflexivity (SR)

$$\{\varphi\} \,\Box\!\!\rightsquigarrow \{\varphi\}$$

An important rule in an aboutness proof system is the Transitivity rule. It states that if $S \,\Box\!\!\rightsquigarrow T$ and $T \,\Box\!\!\rightsquigarrow U$ are concluded, then it is allowed to draw the conclusion that $S \,\Box\!\!\rightsquigarrow U$. If the aboutness decision is based on the existence of some overlap then Transitivity does not hold: an overlap between S and T and between T and U does not imply that there is an overlap between S and U. This kind of decisions occur in vector-space models. For an information-theoretical approach we believe however, that the aboutness property should include this rule.

Transitivity (Tr)

$$\frac{S \,\Box\!\!\rightsquigarrow T \quad T \,\Box\!\!\rightsquigarrow U}{S \,\Box\!\!\rightsquigarrow U}$$

A rule which can cause problems for a number of aboutness theorems is Symmetry. Symmetry expresses the claim that there is no difference between concluding that a situation S is about a situation T and concluding that a situation T is about a situation S.

Symmetry (Sy)

$$\frac{S \,\Box\!\!\rightsquigarrow T}{T \,\Box\!\!\rightsquigarrow S}$$

In some retrieval systems, for example boolean retrieval, Symmetry is precluded by the strict inference mechanism. As we will show in Chapter 4, coordination level matching and vector-space models turn out to be symmetric. The symmetry property is primarily intended to increase the number of aboutness theorems.

Two set-equivalent sets should have the same aboutness decisions. This requirement is modelled with the following Set Equivalence rule:

Set Equivalence (SE)

$$\frac{S \,\Box\!\!\rightsquigarrow U \quad S \equiv T}{T \,\Box\!\!\rightsquigarrow U} \qquad\qquad \frac{S \,\Box\!\!\rightsquigarrow T \quad T \equiv U}{S \,\Box\!\!\rightsquigarrow U}$$

As one can see Set Equivalence contains two rules, namely Left Set Equivalence and Right Set Equivalence. Both rules state that the aboutness derivation between two situations should depend on the meaning of the situations, not on their form. Given that $S \cup T \equiv T \cup S$ holds, we can derive with this rule, for instance that, given $S \cup T \,\Box\!\!\rightsquigarrow S$, it is allowed to conclude that $T \cup S \,\Box\!\!\rightsquigarrow S$.

The last basic rule we present is a good example of a property which should not hold in general within the context of a particular model.

Euclid (Eu)

$$\frac{S \,\Box\!\!\rightsquigarrow T \quad S \,\Box\!\!\rightsquigarrow U}{T \,\Box\!\!\rightsquigarrow U}$$

The rule Euclid expresses that if S is about T and also about U, the conclusion $T \,\Box\!\!\rightsquigarrow U$ can be derived. If an aboutness proof system satisfies this rule, some counterintuitive aboutness derivations can be made.

Combination postulates

The term 'monotonicity', which is frequently used with respect to proof systems in general, stems here from the fact that aboutness is preserved under informational union. An example of the Left Monotonic Union rule is the following. Let the situation $\{\langle\langle \text{Sioux} \rangle\rangle, \langle\langle \text{Cavalry} \rangle\rangle\}$ be about $\{\langle\langle \text{Sioux} \rangle\rangle, \langle\langle \text{Cavalry} \rangle\rangle\}$ (for instance by using Reflexivity), and form a new situation by informationally uniting the first situation with $\{\langle\langle \text{Battle} \rangle\rangle\}$. Left Monotonic Union allows us to conclude that this new situation is also about $\{\langle\langle \text{Sioux} \rangle\rangle, \langle\langle \text{Cavalry} \rangle\rangle\}$.

Left Monotonic Union (LMU)

$$\frac{S \,\Box\!\!\rightsquigarrow T}{S \cup U \,\Box\!\!\rightsquigarrow T}$$

This monotonic rule needs some attention. Adding information leads only to more conclusions, never to a reduction of it. This implicitly means that extending the characterisation of information will possibly make the system decide that there are more aboutness derivations, but never less. At first sight this does not look as an unreasonable property. However, take for example the famous Tweety–Bird example. If we could make the decision that Tweety, being a bird, can fly, then there is no possibility to withdraw this fact by adding information, for example that Tweety is a penguin. In Chapter 5 we prove that if a model is based on only monotonic rules of this kind, then extending the representation of the documents will lead to a better recall. As mentioned in Chapter 2, increasing recall generally leads to decreasing precision. Therefore we have to be careful in adopting this rule without any restrictions.

The Left Monotonic Union can thus be an undesirable postulate. Often user preference plays an important role in the way the addition of information preserves the aboutness relation with the query [27, 28]. The acceptability of a deduction step depends on what the user had in mind. Consider the following example: if a user is interested in "water energy" and gives the query "water", then we can assume that with respect to this particular need, "water mills" is about "water"; nonetheless, the conclusion that "water pollution" is about "water" is not allowed. This kind of user preferences and their non-monotonic behaviour can hardly be generalised. What we can do is formulate some general guarded rules of the following form:

Guarded Left Union (GLU)

$$\frac{S \,\square\!\!\rightsquigarrow T \qquad \mathit{Requirement}}{S \cup U \,\square\!\!\rightsquigarrow T}$$

This is the general formulation of what we call Guarded Left Union. The requirement is to be replaced by a concrete constraint. In the case of Left Monotonic Union the constraint is set to *true* or *void*.

We can now propose a list of possible postulates with a specific substitution for the requirement. For instance the Cautious Monotonicity rule can be proposed. This rule has it origin in the work of Kraus, Lehmann & Magidor [81]. In their work, the authors try to capture the general notion of non-monotonic reasoning.

Cautious Monotonicity (CM)

$$\frac{S \,\square\!\!\rightsquigarrow T \qquad S \,\square\!\!\rightsquigarrow U}{S \cup U \,\square\!\!\rightsquigarrow T}$$

This postulate states that the aboutness relation is not violated by adding to S all the information S is about.

Another suggestion for a Guarded Left Union rule is the following:

Left Related Union (LRU)

$$\frac{S \,\square\!\!\rightsquigarrow T \qquad U \,\square\!\!\rightsquigarrow T}{S \cup U \,\square\!\!\rightsquigarrow T}$$

In this case we have evidence that the situation we unite with S is also about T. Here, we are only extending S with information that is known to be about T. In the next paragraphs we see some more Guarded Left Union rules.

Rather than uniting information at the left side of the aboutness derivation we can propose right variants. For instance, Right Monotonic Union:

Right Monotonic Union (RMU)

$$\frac{S \Box\!\!\rightsquigarrow T}{S \Box\!\!\rightsquigarrow T \cup U}$$

Instead of summing up all the possible variants, we look at a different class of postulates. Rather than adding information we can propose rules that withdraw information as a kind of generalisation. For instance, a situation $\{\langle\langle \text{Sioux} \rangle\rangle\}$ is a generalisation of a situation $\{\langle\langle \text{Sioux} \rangle\rangle, \langle\langle \text{Battle} \rangle\rangle, \langle\langle \text{Cavalry} \rangle\rangle\}$. Similar to the previous Left Monotonic Union rule we can suggest a rule which allows to withdraw any information item of the situation and still keeps the aboutness relation between the two situations valid:

Right Weakening (RW)

$$\frac{S \Box\!\!\rightsquigarrow T \cup U}{S \Box\!\!\rightsquigarrow T}$$

Following the above example, if a situation S is about $\{\langle\langle \text{Sioux} \rangle\rangle, \langle\langle \text{Battle} \rangle\rangle, \langle\langle \text{Cavalry} \rangle\rangle\}$, it is also about $\{\langle\langle \text{Sioux} \rangle\rangle\}$. In this case we can propose the same kind of guards as we proposed with the Guarded Left Union rules.

At this point, we can define similar postulates for the intersection. The intersection of two situations can be viewed as a composition of two situations, and therefore we refer to the following rules as composition rules. Due to their set-theoretical aspects, composition rules have much in common with the union rules. Take for example the right composition:

Composition (Cp)

$$\frac{S \Box\!\!\rightsquigarrow T}{S \Box\!\!\rightsquigarrow T \cap U}$$

The composition property expresses that if a situation S is about a given situation T, aboutness is preserved under any composition of the situation T.

Note that with this rule one can derive that S is about $\emptyset$. If, for some reasons, this type of aboutness theorems should be avoided one could in certain cases[6] adopt the Strict Composition rule instead of the Composition rule. This rule states that if S is about T, then S is about the intersection of S and T.

Strict Composition (SC)

$$\frac{S \Box\!\!\rightsquigarrow T}{S \Box\!\!\rightsquigarrow T \cap S}$$

The following Right Monotonic Decomposition rule clearly represents the idea that adopting this rule it is allowed, given that situation S is about the intersection of situation T and the situation U, to infer that situation S is about situation T.

[6] More precisely, in cases in which the property: if $S \Box\!\!\rightsquigarrow T$ then $S \cap T \not\equiv \emptyset$ holds.

Right Monotonic Decomposition (RMD)

$$\frac{S \,\Box\!\!\rightsquigarrow T \cap U}{S \,\Box\!\!\rightsquigarrow T}$$

For the combining postulates we only highlight one other interesting postulate, the rule Context-Free Union.

Context-Free Union (CFU)

$$\frac{S \,\Box\!\!\rightsquigarrow T \qquad S \,\Box\!\!\rightsquigarrow U}{S \,\Box\!\!\rightsquigarrow T \cup U}$$

This rule states that when one can conclude that $S \,\Box\!\!\rightsquigarrow T$ and $S \,\Box\!\!\rightsquigarrow U$, then one can unite the information of T and U and conclude that S is about this union. Boolean retrieval, for one, is founded on this postulate. For example, if a document d is about $\{\langle\langle\text{Sioux}\rangle\rangle\}$ and the same document is about $\{\langle\langle\text{Cavalry}\rangle\rangle\}$, it is assumed that d is about $\{\langle\langle\text{Sioux}\rangle\rangle, \langle\langle\text{Cavalry}\rangle\rangle\}$. In this case we have to be sure that we are not able to draw more information from the situation $\{\langle\langle\text{Sioux}\rangle\rangle, \langle\langle\text{Cavalry}\rangle\rangle\}$ than the fact that both keywords are present in the document. It is for instance not allowed to assume that there exists a relation between "Sioux" and "cavalry".

The Cut rule is common in logical systems in order to extend the deduction possibilities:

Cut (Cu)

$$\frac{S \cup T \,\Box\!\!\rightsquigarrow U \quad S \,\Box\!\!\rightsquigarrow T}{S \,\Box\!\!\rightsquigarrow U}$$

Assume that we want to prove that S is about U and we already know that S is about T. Adopting this rule implies that for obtaining the derived conclusion it is enough to prove that $S \cup T$ is about U.

Infon-based postulates

So far we proposed postulates without taking into account the properties of the infons. However the relation $\rightarrow$ (information containment) may be useful for the aboutness decision. The following rule claims that all rules valid for the information containment are valid for the information aboutness:

Containment (Cm)

$$\frac{\varphi \rightarrow \psi}{\{\varphi\} \,\Box\!\!\rightsquigarrow \{\psi\}}$$

We can propose an extra premise of the Containment rule in order to extent the situations which are about each other.

Union Containment (UC)

$$\frac{\varphi{\to}\psi \qquad S \cup \{\psi\} \,\Box\!\!\rightsquigarrow T}{S \cup \{\varphi\} \,\Box\!\!\rightsquigarrow T}$$

This rule expresses the fact that if situation S' is about situation T, with $\psi \in S'$ and it is given that $\varphi{\to}\psi$, then we can 'replace' the infon ψ with φ in S' without loosing its being about situation T.

The use of the preclusion relation between infons can be proposed for a restricted version of the monotonic union. For instance, if it can be established that if a situation S is about a situation T, then adding an infon φ to the situation S and an infon ψ to the situation T will not violate this, provided no preclusion between φ and ψ is apparent.

Compositional Monotonicity (CM)

$$\frac{S \,\Box\!\!\rightsquigarrow T \qquad \varphi \not\perp \psi}{S \cup \{\varphi\} \,\Box\!\!\rightsquigarrow T \cup \{\psi\}}$$

Obviously we can use the information containment and preclusion relations as a guard for the combination postulates. For instance, in the case of Guarded Left Union rules, the guard restricts the situation that could be added to the situation on the left side of the aboutness. Guarded Union Containment, for example, states that it is only possible to add infons (representing singleton situations) which are informationally contained in the original situation on the left side. Unfortunately our language does not have the ability to express the requirement that $\varphi \in S$, in order to express that $\varphi{\to}\psi$ for some $\varphi \in S$. Therefore we suggest:

Guarded Union Containment (GUC)

$$\frac{\varphi{\to}\psi \qquad S \cup \{\varphi\} \,\Box\!\!\rightsquigarrow T}{S \cup \{\varphi\} \cup \{\psi\} \,\Box\!\!\rightsquigarrow T}$$

We can also adopt rules to axiomatise the notion of preclusion. For instance the following rule is proposed by Seligman [135]:

Mutual Preclusion (MP)

$$\frac{\varphi \perp \psi}{\psi \perp \varphi}$$

Seligman noticed that the meaning of preclusion in English is slightly different. However, adopting this rule in order to propose a symmetric preclusion would not be problematic.

The relation of certain infons can also play a role in the aboutness derivation. For instance, the logical connective $\vee$ of the boolean infon language may be important. Here we could propose a rule stating that, if $S \,\Box\!\!\rightsquigarrow \{\varphi\}$ then also $S \,\Box\!\!\rightsquigarrow \{\langle\langle \vee, \varphi, \psi;\ 1\rangle\rangle\}$.

R-Right Monotonic Composition (R-RMC)

$$\frac{S \,\Box\!\!\leadsto \{\varphi\}}{S \,\Box\!\!\leadsto \{\langle\langle R, \varphi, \psi;\ 1\rangle\rangle\}}$$

Thus, in case the rule involves the connective $\vee$ the rule is called the $\vee$-Right Monotonic Composition rule.

3.3.2 Reasoning with situation anti-aboutness

As mentioned in Chapter 2, it is useful to study the cases in which a theory is refuted. Or stated differently, when aboutness should not be derived. For this reason we introduce the anti-aboutness relation. As we explained with Example 2.3 at page 28, a model that returns 'Planes with wings' given a query 'Flying objects without wings' could be improved if we have properly defined the notion of anti-aboutness.

In our opinion, an IR model should not only be good in determining aboutness, but also in distinguishing the anti-aboutness relations. To forestall confusion, the anti-aboutness relation expresses that two situations are each other's opposite and *not* that two situations are *not* about each other. As we believe that these two relations are not equivalent it should be noted that, if we are not able to prove aboutness, this does not imply that we proved anti-aboutness.

Example 3.2 Take for example the following three situations:

$$\begin{aligned} S &= \{\langle\langle \textit{defeated}, \langle\langle \text{Sioux}\rangle\rangle, \langle\langle \text{Cavalry}\rangle\rangle;\ 1\rangle\rangle\} \\ T &= \{\langle\langle \textit{defeated}, \langle\langle \text{Sioux}\rangle\rangle, \langle\langle \text{Cavalry}\rangle\rangle;\ 0\rangle\rangle\} \\ U &= \{\langle\langle \textit{killed}, \langle\langle \text{Brutus}\rangle\rangle, \langle\langle \text{Caesar}\rangle\rangle;\ 1\rangle\rangle\} \end{aligned}$$

If one is interested in a situation $V = \{\langle\langle \textit{defeated}, \langle\langle \text{Sioux}\rangle\rangle, \langle\langle \text{Cavalry}\rangle\rangle;\ 1\rangle\rangle\}$, we can intuitively construct the following table:

	about V	anti-about V
S	Yes	No
T	No	Yes
U	No	No

Now, there is a difference between the situations T and U with respect to the situation V. An IR model that considers the situations S and U as being about situation V should in our opinion be preferred over an IR model that considers the situations S and T to be about situation V, since the errors of the retrieved set of the latter model will be more confusing to the user than the errors of the former. Any user would recognise the document corresponding to situation U immediately as not relevant and therefore this document is more harmless than if the document corresponding to situation T is showed to the user. In this case, she needs to inspect it more closely in order to recognise that the retrieved document is *definitely* not what she was interested in.

So, we propose a new relation for our language. The relation *anti-aboutness* (denoted by $\boxtimes\!\!\leadsto$) expresses the fact that a situation S is in conflict with a situation T, denoted by $S\boxtimes\!\!\leadsto T$.

Definition 3.11 For a given infon language $\mathcal{I}$, the *extended aboutness language* $\mathcal{L}^{Ext}(\mathcal{I})$ of aboutness formulae is the smallest superset of the aboutness language $\mathcal{L}(\mathcal{I})$ such that

- if $S, T \in \mathcal{S}(\mathcal{I})$ then $S\boxtimes\!\!\leadsto T, S\boxtimes\!\!\not\leadsto T \in \mathcal{L}^{Ext}(\mathcal{I})$

where $\mathcal{L}(\mathcal{I})$ and $\mathcal{S}(\mathcal{I})$ are as in Definition 3.9.

Whenever the language $\mathcal{I}$ is understood we write $\mathcal{L}^{Ext}$ rather than $\mathcal{L}^{Ext}(\mathcal{I})$.

One possibly desirable property will be that it is impossible to deduce aboutness and anti-aboutness at the same time. An aboutness proof system that precludes such possibilities is termed *consistent*:

Definition 3.12 An aboutness proof system $\mathcal{A}_{ps} = \langle \mathcal{L}^{Ext}, Ax, Rule \rangle$ is called *consistent* if and only if there are no S and T in $\mathcal{S}$ such that

$$\vdash_{\mathcal{A}_{ps}} S \,\Box\!\!\leadsto T \text{ and } \vdash_{\mathcal{A}_{ps}} S\boxtimes\!\!\leadsto T.$$

In case an aboutness proof system is not consistent, it is termed *inconsistent*.

If one wants to adopt a simple definition of anti-aboutness the following rule could be adopted:

Simple Anti-Aboutness (SAA)

$$\frac{S \,\Box\!\!\not\leadsto T}{S\boxtimes\!\!\leadsto T}$$

Here the, in our opinion wrong, assumption has been made that if we are not able to prove aboutness this implies that we proved anti-aboutness.

Example 3.2 showed us that the preclusion operator can be of great help for determining anti-aboutness:

Preclusion (Pr)

$$\frac{\varphi \perp \psi}{\{\varphi\}\boxtimes\!\!\leadsto\{\psi\}}$$

If information items preclude each other, then it does not seem unreasonable to assume that the situations that support only these particular infons are anti-about each other. Applications of this assumption can be readily found in information retrieval. Another use of preclusion as a premise of a rule that we would like to mention explicitly is the one Seligman termed *local preclusion* [135]. In terms of our framework the rule could be presented as:

Local Preclusion (LP)

$$\frac{S\,\Box\!\!\rightsquigarrow\{\varphi\} \quad \varphi\perp\psi}{S\boxtimes\!\!\rightsquigarrow\{\psi\}}$$

If a situation is anti-about $\{\langle\langle\text{Sioux}\rangle\rangle\}$, then it is likely to assume that the situation is anti-about$\{\langle\langle\text{Sioux}\rangle\rangle, \langle\langle\text{Cavalry}\rangle\rangle\}$ (as we are already convinced that the information of S is anti-about the "Sioux", how could the information of S be about the "Sioux" and "cavalry"?). This is the intuition behind the so-called Negation Rationale.

Negation Rationale (NR)

$$\frac{S\boxtimes\!\!\rightsquigarrow T}{S\boxtimes\!\!\rightsquigarrow T\cup U}$$

If a situation is anti-about another situation, then no aboutness relation can be established by adding information to the conclusion. This postulate stands in close relation with the non-monotonicity behaviour. In order to create only consistent aboutness proof systems, we have to be careful to adopt this rule together with the Right Monotonic Union rule. For instance, the following proposition presents an inconsistent aboutness proof system.

Proposition 3.1 The aboutness proof system $\mathcal{A}_{ps}$ defined by $\langle\mathcal{L}^{Ext}, \{S\,\Box\!\!\rightsquigarrow U, S\boxtimes\!\!\rightsquigarrow T\},$ $\{\text{Set Equivalence}, \text{Right Monotonic Union}, \text{Negation Rationale}\}\rangle$ is inconsistent.

Proof Given the axiom $S\,\Box\!\!\rightsquigarrow U$, using the Right Monotonic Union rule, we can derive $S\,\Box\!\!\rightsquigarrow T\cup U$. At the same time, given the axiom $S\boxtimes\!\!\rightsquigarrow T$, applying the rule Negation Rationale the formula $S\boxtimes\!\!\rightsquigarrow U\cup T$ can be derived. With the Set Equivalence rule, we can determine that $S\,\Box\!\!\rightsquigarrow U\cup T$ and $S\boxtimes\!\!\rightsquigarrow U\cup T$, which proves the inconsistency. □

In order to avoid such kind of problems, we introduce the rule Cautious Negation Rationale, which is formulated as follows:

Cautious Negation Rationale (CNR)

$$\frac{S\boxtimes\!\!\rightsquigarrow T \quad S\,\Box\!\!\not\rightsquigarrow U}{S\boxtimes\!\!\rightsquigarrow T\cup U}$$

Given that situation S is anti-about situation T, this rule only allows us to extend the anti-aboutness conclusion to the fact that situation S is anti-about situation $T\cup U$ if it is not possible to prove that the situation S is about U.

3.3.3 Combining aboutness proof systems

As mentioned in Chapter 1, the information retrieval problem also concerns the problem of having information available at different places. In the *old information retrieval* paradigm there was a one-to-one correspondence between a user and a document-collection. Nowadays, the information retrieval problem is a matter of a many-to-many relation, as different kinds of users are searching for information in different information domains, possibly stretched out over the globe.

This brings in a new requirement which concerns the possibility of combining different IR models in order to create a new IR model. A combination can only be justified if we have a deep insight in the effectiveness of the proposed combination. In this section we present a study of the combination of aboutness proof systems in terms of the framework.

In order to formalise an aboutness proof system that can use different aboutness-notions, a kind of *combined aboutness language* is needed. This language should present a combination of several different aboutness languages. We use the relation X_i to refer to the relation X of aboutness language i for all $X \in \{\rightarrow, \perp, \not\rightarrow, \not\perp, \Box\!\!\leadsto, \Box\!\!\not\leadsto\}$. For instance, combining two aboutness languages we may have the two aboutness relations $\Box\!\!\leadsto_1$ and $\Box\!\!\leadsto_2$ as elements of the combined aboutness language.

A more formal definition of a *combined aboutness language* is given as follows:

Definition 3.13 For a given infon language $\mathcal{I}$, the *combined aboutness language* $\mathcal{L}_n(\mathcal{I})$ of aboutness formulae is the smallest set such that

- if $\varphi, \psi \in \mathcal{I}$ then $\varphi \rightarrow_i \psi$,$\varphi \perp_i \psi$,$\psi \not\rightarrow_i \varphi$, $\varphi \not\perp_i \psi \in \mathcal{L}_n(\mathcal{I})$
- if $S, T \in \mathcal{S}(\mathcal{I})$ then $S \Box\!\!\leadsto_i T, S \Box\!\!\not\leadsto_i T, S \equiv_i T, S \not\equiv_i T \in \mathcal{L}_n(\mathcal{I})$

where $\mathcal{S}(\mathcal{I})$ is as in Definition 3.8 and $i \in \mathbb{N}$ with $1 \leq i \leq n$.

Whenever the language $\mathcal{I}$ is understood we write $\mathcal{L}_n$ rather than $\mathcal{L}_n(\mathcal{I})$. For the sake of brevity, we will not introduce the *extended combined aboutness language* $\mathcal{L}_n^{Ext}(\mathcal{I})$ (or $\mathcal{L}_n^{Ext}$ for short); its definition proceeds in a similar way.

An aboutness proof system using the language $\mathcal{L}_n$ (or $\mathcal{L}_n^{Ext}$) presents n different notions of aboutness (respectively notions of anti-aboutness). We will term such a proof system a *combined aboutness proof system.* Theorems of the form $S \Box\!\!\leadsto_k T$ are called *Aboutness Theorems of k.*

The first rule in an extended language one may think of is that all aboutness theorems of i are aboutness theorems of j. This aboutness meta-property is included in the following rule.

Aboutness Inheritance (AI)

$$\frac{S \Box\!\!\leadsto_i T}{S \Box\!\!\leadsto_j T}$$

Another typical example of a rule in a language $\mathcal{L}_n^{Ext}$ is the *Closed World Assumption* as introduced by Reiter [120]:

Closed World Assumption (CWA)

$$\frac{S \boxtimes\!\!\leadsto_i \{\varphi\} \quad \varphi \perp_j \psi}{S \;\Box\!\!\leadsto_j \{\psi\}}$$

Given that S is anti-about $\{\varphi\}$ in terms of aboutness proof system i and in aboutness proof system j the axiom $\varphi \perp \psi$ is present, it follows that S is about $\{\psi\}$ in proof system j. This rule is typically used in databases. Normally, in databases only positive facts are recorded. If a fact is not recorded, one may assume that the opposite of the fact holds. As we will see in Chapter 4, some IR models adopt this rule (see also [67, 71]). Using the same arguments that were used to question the definition of anti-aboutness in terms of not about, the usefulness of this CWA rule can be questioned.

3.3.4 Information retrieval agents

Instead of proposing new rules, we can use a combined aboutness proof system to formalise the concepts of the new information retrieval paradigm as presented at page 8. In this paradigm, there is no *general* search strategy, due to different kinds of users and different kinds of search-actions (e.g., searching for general information or for detailed information). Another point is that a search for information consists of different kinds of searches on a broad range of information collections. It is not surprising that when trying to meet these new requirements one looks at approaches that have taken root in AI. The motivation that rational agents can be used as atomic IR systems, with the ability to reason, communicate, and gather information is proposed by several authors [33, 70, 91].

Van Linder [90] has recently presented a first attempt at a formalisation of information agents. Our approach is based on similar ideas. The formalisation in modal logic [70, 72] is beyond the scope of this thesis.

Based on the intuitive ideas in the thesis of Van Linder, we consider rational agents with the ability to reason, communicate, and gather information. We recognise two types of agents, the *retrievers* and the *users*. The retriever agents decide whether a document is about (or anti-about) a query. Since we are not adopting a Closed World definition, where anti-aboutness would simply be defined to be the absence of aboutness, the retriever agents can decide whether a document representation is anti-about to a query.

In terms of our framework we can easily formulate an agent as an element of a combined proof system. These agents have the ability to conclude whether a document is about (or anti-about) a query.

Given an aboutness proof system $\mathcal{A}_{ps} = \langle \mathcal{L}_n^{Ext}, Ax, Rule \rangle$, we define k retriever agents and l user agents with $k + l = n$. Each retriever agent represents a unique concept of aboutness, for instance the one of a vector-space model or a boolean model. For the rest of this section let us assume that we have defined k retriever agents, denoted as r_i with $1 \leq i \leq k$, and a corresponding aboutness decision $\Box\!\!\leadsto_{r_i}$ and anti-aboutness decision $\boxtimes\!\!\leadsto_{r_i}$ for each of them.

The communication and information gathering is done by the so-called user agents. In reality different users have different concepts of aboutness. We can also distinguish different kinds of user agents in the way they conclude aboutness given the retrieval results obtained by the retriever agents.

The first user u_1 we can think of is a typical user, who is satisfied with a document if at least one retriever states that the document is about the query.

Typical User (TU)

$$\frac{S \Box\!\!\leadsto_{r_1} T}{S \Box\!\!\leadsto_{u_1} T} \quad \cdots \quad \frac{S \Box\!\!\leadsto_{r_k} T}{S \Box\!\!\leadsto_{u_1} T}$$

The second type of user u_2 is more like a lawyer, who is preparing a case, and therefore considers a document about a query if none of the retrievers consider the document anti-about the query. So, this user is sure not to forget some material which could possibly be relevant.

Lawyer (La)

$$\frac{S \boxtimes\!\!\not\leadsto_{r_1} T \quad \ldots \quad S \boxtimes\!\!\not\leadsto_{r_k} T}{S \Box\!\!\leadsto_{u_2} T}$$

The third user u_3 is a more careful one, and considers a document to be about a query if one of the retriever agents considers it about the query while the others do not consider it to be anti-about the query.

Careful User (CU)

$$\frac{S \boxtimes\!\!\not\leadsto_{r_1} T \quad \ldots \quad S \boxtimes\!\!\not\leadsto_{r_{i-1}} T \quad S \Box\!\!\leadsto_{r_i} T \quad S \boxtimes\!\!\not\leadsto_{r_{i+1}} T \quad \ldots \quad S \boxtimes\!\!\not\leadsto_{r_k} T}{S \Box\!\!\leadsto_{u_3} T}$$

The last user u_4 we present is a very careful one, which is satisfied with a document if all retriever agents state that the document is about the query. So, this user requires a complete agreement upon the aboutness decision.

Unanimous User (UU)

$$\frac{S \Box\!\!\leadsto_{r_1} T \quad \ldots \quad S \Box\!\!\leadsto_{r_k} T}{S \Box\!\!\leadsto_{u_4} T}$$

Some possible properties of the retriever agents with respect to the user agents can be defined.

Definition 3.14 A combined aboutness proof system $\mathcal{A}_{ps}$ is called *empty* with respect to a user agent u_i if there are no situations S and $T \in \mathcal{S}$ such that $\vdash_{\mathcal{A}_{ps}} S \,\Box\!\!\leadsto_{u_i} T$.

Definition 3.15 A combined aboutness proof system $\mathcal{A}_{ps}$ is called *overfull* with respect to a user agent u_i if for all situations S and $T \in \mathcal{S}$ holds that $\vdash_{\mathcal{A}_{ps}} S \,\Box\!\!\leadsto_{u_i} T$.

Note that, the aboutness decision of an unanimous user is more strict in the sense of aboutness than the one of a typical user. If we know that a document d_1 is about query q for an unanimous user and d_2 is about query q for a typical user and not for an unanimous user, we can draw the conclusion that we prefer document d_1 over document d_2 with respect to the query q. Chapter 6 presents an ordering method for aboutness proof systems to come to a *ranking* of documents based on above motivations.

This section has presented a first approach towards a combination of different IR models based on qualitative grounds. The combination is not based on recall and precision values but on the derivation aspects of different models.

3.4 Summary and conclusions

In this chapter we have presented a framework for information retrieval based on an underlying theory of information. Within this framework, formal representatives of documents and their characterisation can be formulated. By proposing a set of postulates, the implicit assumptions governing an information retrieval mechanism can be brought to light. The effectiveness of a retrieval mechanism can be examined, not only by running experiments, but by inspecting the postulates of the model. In the next chapter we investigate theoretical and existing IR models using this theory.

Chapter 4

IR models and their aboutness proof systems

I cannot refute you, Socrates, said Agathon:
– Let us assume that what you say is true.
Say rather, beloved Agathon, that you cannot
refute the truth; for Socrates is easily refuted.

Plato, *'Symposium'*.

In this chapter we investigate different IR models, in order to explore the strength of some general assumptions of aboutness. We start with a very basic IR model, based on a subset relation, and end with some logical IR models. Using the framework introduced in Chapter 3, we look at the aboutness properties that characterise these IR models.

We have motivated that aboutness is the formal counterpart of relevance, based on the relevance definition of Cooper [41]. Van Rijsbergen [123, 124] proposed in 1986 that an *unspecified* non-classical conditional logic should be used to deduce aboutness. In his framework aboutness decisions are interpreted through logical inference: a document d is about a query q if q can be proved from d. If q cannot be proved from d, however, then no definitive statement can be made about d being about q.

After Van Rijsbergen's proposal, the logical approach to information retrieval has gained quite some attention. A wide range of logical IR models were proposed based on, for example, Modal Logic [106, 109, 107], Conceptual Graphs [34, 78, 79, 99], Refinement Machines [23, 24], Terminological Logic [100, 101, 134], Abductive Logic [103, 104], Datalog [55], Logical Imaging [45, 46], and Situation Theory [84, 86, 87, 126, 127]. All these proposals are part of the quest for the *one and only* logic for information retrieval.

Another use of logic for information retrieval is presented by Chiaramella & Chevallet in their article 'About Retrieval Models and Logic' [35]. In this article, the authors are not proposing a *new* logic for information retrieval. They are using logic as a vehicle to analyse what they call 'some lesser known aspects of information retrieval, as for example the impact of new applications [such as multi-media] which already induce a complete revision of the notion of document'.

In general terms they discuss in which way the semantics of the logical inference, the semantical content of a document, the use of contextual attributes, and the semantical representation of a query can be presented. By inspecting a typical logical IR model, namely the boolean model, the authors make clear that one can improve a logical IR model based on assumptions of what an IR logic should be. Furthermore they showed that the benefits of using logic for such an investigation lies in the expressive power, or generality, of logic on the one side, and its very close relation with the fundamentals of information retrieval on the other. This logic-based approach to information retrieval is adopted by several authors [84, 85, 107, 112, 126].

The approach presented in this chapter is, like that of Chiaramella & Chevallet, a study of information retrieval on a meta-level. It differs from the original logical framework of Van Rijsbergen and Chiaramella & Chevallet in that we do not start with a *begin* situation (referring to the document) and conclude via logical deduction steps an *end* situation (referring to the query). We start with axioms stating *what is about what*, and conclude via deduction steps whether a situation (referring to the document) is about another situation (referring to the query).

Our approach is based on the work of Kraus, Lehman & Magidor [81], who present a general framework in which non-monotonic inferences of logical systems can be compared and classified. Their study of the inference-relation concentrates on properties that are or should be enjoyed by non-monotonic reasoning systems, and their meta-theory has by now become the standard one for characterising non-monotonic inference relations. The main interest of Kraus, Lehman & Magidor was to study non-monotonic relations, not to propose a new model:

> *'The different families of models described in this paper and that provide semantics to the axiomatic systems are not considered to be an ontological justification for our interest in the formal systems, but only as a technical tool to study those systems and in particular settle questions of interderivability and find efficient decisions procedures.'*

Our study of the concept of aboutness is based on the same grounds. We do not propose our framework with the intention to create a new (formal) model, but do it to study different IR systems. Hence, our study is focused on the properties of the aboutness relation that are or should be enjoyed by IR systems. Therefore we will analyse several common IR systems using the general framework presented in Chapter 3. For each system we axiomatise the aboutness decisions.

Investigation of an IR model

The study of an IR model proceeds as follows. We present an IR model $\mathcal{A}_m$ and its aboutness decision, denoted as $\models_{\mathcal{A}_m} d$ about q. Here, we formulate the statement 'd about q if and only if...' in terms of the model. Next, the representation of the model as an aboutness proof system is given. In terms of our framework, we present an aboutness proof system $\mathcal{A}_{ps}$ that derives aboutness decisions, denoted as $\vdash_{\mathcal{A}_{ps}} S \,\Box\!\!\leadsto T$. In order to translate document representations and queries into situations, we use a function map that maps document representations $\chi(d)$ and queries q to situations. Here, we define a function that maps a representation language onto another representation language. In case the representation language of the document representation and the one of the query differs we define the functions map_1 and map_2, but we assume that this is not done unless explicitly stated otherwise. After every introduction, we inspect the function map to see whether it is injective, surjective, or bijective.

Axioms and rules are distilled from the properties of the given IR model. To this end, we use the framework defined in Chapter 3. The IR model is presented in terms of axioms and rules of an aboutness proof system $\mathcal{A}_{ps}$ where a document d is about a query q if and only if $map(\chi(d)) \,\Box\!\!\leadsto map(q)$ can be proved using the proof system, or stated differently, if $map(\chi(d)) \,\Box\!\!\leadsto map(q)$ is an *aboutness theorem*.

We inspect the connection between the derivation of aboutness decisions $\vdash_{\mathcal{A}_{ps}} S \,\Box\!\!\leadsto T$ with our proposed logic and the given model aboutness decision $\models_{\mathcal{A}_m} d$ about q. In logic, the properties of the relation between the model and the proof system are formalised in terms of *soundness* and *completeness* theorems. The soundness theorem assures us that the restrictions of the rules are sufficient to block all undesirable conclusions that might otherwise be drawn. The completeness theorem assures us that the rules are in themselves sufficient to generate all valid argument schemata; nothing has been forgotten [57].

In the case of (predicate) logic, soundness and completeness deal with the connection between the inference rules (syntax) and validity in certain models (semantics). In our theory the connection is between the logical deduction of aboutness $\vdash_{\mathcal{A}_{ps}} S \,\Box\!\!\leadsto T$, a syntactic approach to aboutness, and the aboutness notion of an IR model $\models_{\mathcal{A}_m} d$ about q, which can be viewed as the semantic approach to aboutness [66].

In order to prove that an aboutness proof system is sound with respect to an IR model, it suffices to prove that the following two requirements are satisfied by the proof system. First, we have to show that each axiom of the aboutness proof system is sound, e.g., is indeed an aboutness decision of the IR model. Second, all its rules should be sound. This allows us to conclude by induction on the length of the derivation that the aboutness proof system is sound.

In logic a truth preserving proof system is required, i.e., truth and falsity in the proof system should correspond to truth and falsity in the model. Here, we would like to have an *aboutness preserving* proof system, i.e., the notion of 'aboutness' in the proof system should correspond to the notion of 'aboutness' in the IR model. In case the aboutness proof system is sound, then every aboutness theorem deducible by the proposed aboutness proof system is indeed an aboutness decision of the model.

For the proof of completeness we have to show that every aboutness decision of the IR model is always an aboutness theorem of the aboutness proof system. In case the proof system is complete, every aboutness decision made by the model is deducible as aboutness theorem in the proposed proof system.

Given a sound and complete aboutness proof system, we can use a function $answer$ that maps an aboutness proof system $\mathcal{A}_{ps}$, a query q, and a document-base $\mathcal{D}$ into a set of documents (a subset of the document-base $\mathcal{D}$). This function is defined by $answer(\mathcal{A}_{ps}, q, \mathcal{D}) =^{def} \{d \in \mathcal{D} \mid \vdash_{\mathcal{A}_{ps}} map(\chi(d)) \Box\!\!\rightsquigarrow map(q)\}$ where $map(\chi(d))$ is the situation representation of the descriptor set $\chi(d)$ of a document d and $map(q)$ the situation representation of a query q.

There are some typical elements in the IR model as well as in the aboutness proof system. Here we define four possible typical elements of an aboutness proof system:

Definition 4.1 Let $\mathcal{A}_{ps}$ be an aboutness proof system and $\mathcal{S}$ a language of situations.

The top query of $\mathcal{A}_{ps}$ (denoted by $\mathbf{1}^q_{\mathcal{A}_{ps}}$) is a (possibly empty) subset of $\mathcal{S}$ and is defined by:

$$\mathbf{1}^q_{\mathcal{A}_{ps}} =^{def} \{T \mid \text{for all } S \in \mathcal{S} \ \vdash_{\mathcal{A}_{ps}} S \Box\!\!\rightsquigarrow T\}.$$

The bottom query of $\mathcal{A}_{ps}$ (denoted by $\mathbf{0}^q_{\mathcal{A}_{ps}}$) is a (possibly empty) subset of $\mathcal{S}$ and is defined by:

$$\mathbf{0}^q_{\mathcal{A}_{ps}} =^{def} \{T \mid \text{for all } S \in \mathcal{S} \ \not\vdash_{\mathcal{A}_{ps}} S \Box\!\!\rightsquigarrow T\}.$$

The top document of $\mathcal{A}_{ps}$ (denoted by $\mathbf{1}^d_{\mathcal{A}_{ps}}$) is a (possibly empty) subset of $\mathcal{S}$ and is defined by:

$$\mathbf{1}^d_{\mathcal{A}_{ps}} =^{def} \{S \mid \text{for all } T \in \mathcal{S} \ \vdash_{\mathcal{A}_{ps}} S \Box\!\!\rightsquigarrow T\}.$$

The bottom document of $\mathcal{A}_{ps}$ (denoted by $\mathbf{0}^d_{\mathcal{A}_{ps}}$) is a (possibly empty) subset of $\mathcal{S}$ and is defined by:

$$\mathbf{0}^d_{\mathcal{A}_{ps}} =^{def} \{S \mid \text{for all } T \in \mathcal{S} \ \not\vdash_{\mathcal{A}_{ps}} S \Box\!\!\rightsquigarrow T\}.$$

If $\mathcal{A}_{ps}$ is a sound and complete aboutness proof system of an IR model, then each document descriptor for which $map(\chi(d)) \in \mathbf{1}^d_{\mathcal{A}_{ps}}$ will always be retrieved. Indeed, every typical element of $\mathcal{A}_{ps}$ has a counterpart in the IR model.

Definition 4.2 Let $\mathcal{A}_{ps}$ be a sound and complete aboutness proof system of an IR model $\mathcal{A}_m$, $\mathcal{D}$ a document-base and q a query.

The top query of $\mathcal{A}_m$ (denoted by $\mathbf{1}^q_{\mathcal{A}_m}$) is a (possible empty) set of queries and is defined by:

$$\mathbf{1}^q_{\mathcal{A}_m} =^{def} \{q \mid answer(\mathcal{A}_{ps}, q, \mathcal{D}) = \mathcal{D}\}.$$

The bottom query of $\mathcal{A}_m$ (denoted by $\mathbf{0}^q_{\mathcal{A}_m}$) is a (possible empty) set of queries and is defined by:

$$\mathbf{0}^q_{\mathcal{A}_m} =^{def} \{q \mid answer(\mathcal{A}_{ps}, q, \mathcal{D}) = \emptyset\}.$$

The top document of $\mathcal{A}_m$ (denoted by $\mathbf{1}^d_{\mathcal{A}_m}$) is a (possible empty) set of documents of $\mathcal{D}$ and is defined by:

$$\mathbf{1}^d_{\mathcal{A}_m} =^{def} \{d \mid \text{for all } q \;\; d \in answer(\mathcal{A}_{ps}, q, \mathcal{D})\}.$$

The bottom document of $\mathcal{A}_m$ (denoted by $\mathbf{0}^d_{\mathcal{A}_m}$) is a (possible empty) set of documents of $\mathcal{D}$ and is defined by:

$$\mathbf{0}^d_{\mathcal{A}_m} =^{def} \{d \mid \text{for all } q \;\; d \notin answer(\mathcal{A}_{ps}, q, \mathcal{D})\}.$$

Note that the top query of $\mathcal{A}_m$ represents those queries for which each document is about. The bottom query of $\mathcal{A}_m$ represent those queries for which no document is about. Those documents that are always retrieved no matter what the query is are elements of the top document of $\mathcal{A}_m$. Finally, the bottom document of $\mathcal{A}_m$ represents those documents that are never retrieved no matter what the query is.

It is enlightening to study the nature of the relationship between the notions given above and the underlying IR model in order to present differences with other models. One may, for instance, wonder whether $\mathbf{1}^d_{\mathcal{A}_{ps}} = \{map(\chi(d)) \mid d \in \mathbf{1}^d_{\mathcal{A}_m}\}$ holds. The following proposition sheds some light on this issue.

Proposition 4.1 Let $\mathcal{A}_{ps}$ be a sound and complete aboutness proof system of an IR model $\mathcal{A}_m$, $\mathcal{D}$ a document-base and q a query. Furthermore, let map be a surjective function. Then

$$\begin{array}{lcllcl}
\mathbf{1}^q_{\mathcal{A}_{ps}} & = & \{map(q) \mid q \in \mathbf{1}^q_{\mathcal{A}_m}\} & \mathbf{1}^q_{\mathcal{A}_m} & = & \{q \mid map(q) \in \mathbf{1}^q_{\mathcal{A}_{ps}}\} \\
\mathbf{0}^q_{\mathcal{A}_{ps}} & = & \{map(q) \mid q \in \mathbf{0}^q_{\mathcal{A}_m}\} & \mathbf{0}^q_{\mathcal{A}_m} & = & \{q \mid map(q) \in \mathbf{0}^q_{\mathcal{A}_{ps}}\} \\
\mathbf{1}^d_{\mathcal{A}_{ps}} & = & \{map(\chi(d)) \mid d \in \mathbf{1}^d_{\mathcal{A}_m}\} & \mathbf{1}^d_{\mathcal{A}_m} & = & \{d \mid map(\chi(d)) \in \mathbf{1}^d_{\mathcal{A}_{ps}}\} \\
\mathbf{0}^d_{\mathcal{A}_{ps}} & = & \{map(\chi(d)) \mid d \in \mathbf{0}^d_{\mathcal{A}_m}\} & \mathbf{0}^d_{\mathcal{A}_m} & = & \{d \mid map(\chi(d)) \in \mathbf{0}^d_{\mathcal{A}_{ps}}\}
\end{array}$$

Proof The proposition follows, for each item, directly from the fact that $\mathcal{A}_{ps}$ is a sound and complete aboutness proof system of a model $\mathcal{A}_m$. So, if $\vdash_{\mathcal{A}_{ps}} map(\chi(d)) \Box\!\!\rightsquigarrow map(q)$ then $\models_{\mathcal{A}_m} d$ about q and vice versa. Note that the requirement of surjectivity of the map function is indeed necessary. For in case map is not surjective, we could have that $\mathbf{1}^d_{\mathcal{A}_{ps}} = \{S\}$ while for all $d \in \mathcal{D}$ $map(\chi(d)) \neq S$. Assume towards a contradiction that $\mathbf{1}^d_{\mathcal{A}_{ps}} = \{map(\chi(d)) \mid d \in \mathbf{1}^d_{\mathcal{A}_m}\}$ and $\mathbf{1}^d_{\mathcal{A}_m} = \{d \mid map(\chi(d)) \in \mathbf{1}^d_{\mathcal{A}_{ps}}\}$. Since for all $d \in \mathcal{D}$ $map(\chi(d)) \neq S$, it follows that $\mathbf{1}^d_{\mathcal{A}_m} = \emptyset$. But this implies that $\mathbf{1}^d_{\mathcal{A}_{ps}} = \emptyset$, which contradicts the assumption that $\mathbf{1}^d_{\mathcal{A}_{ps}} = \{S\}$.

□

Note, that we do not require the function map to be injective. In case the function map is not injective the proposition still holds. A non-injective function map allows that $map(x) = map(y)$ with the possibility that $x \neq y$. However, if $S \in \mathbf{1}^d_{\mathcal{A}_{ps}}$ and $S = map(d_1) = map(d_2)$ then by definition d_1 and $d_2 \in \mathbf{1}^d_{\mathcal{A}_m}$. Conversely, if $d_1, d_2 \in \mathbf{1}^d_{\mathcal{A}_m}$, then $map(d_1) \in \mathbf{1}^d_{\mathcal{A}_{ps}}$ and $map(d_2) \in \mathbf{1}^d_{\mathcal{A}_{ps}}$, thus $S \in \mathbf{1}^d_{\mathcal{A}_{ps}}$.

For each model and system we can represent the top and bottom sets. However as Proposition 4.2 states, it is possible to state that some sets are empty based on the fact that other sets are not empty.

Proposition 4.2 Let $X \in \{\mathcal{A}_{ps}, \mathcal{A}_m\}$, $I \in \{\mathbf{0}, \mathbf{1}\}$, and $\alpha \in \{q, d\}$. Define the function $\overline{x}$ to be the complement function, i.e., $\overline{\mathbf{0}} = \mathbf{1}$ and $\overline{\mathbf{1}} = \mathbf{0}$, and $\overline{q} = d$ and $\overline{d} = q$. Then

if $I^{\alpha}_{\mathrm{X}} \neq \emptyset$ then $\overline{I}^{\overline{\alpha}}_{\mathrm{X}} = \emptyset$.

Proof We show the proposition for $X = \mathcal{A}_{ps}$, $I = \mathbf{1}$ and $\alpha = d$; the other cases are proved analogously. We have to show that if $\mathbf{1}^d_{\mathcal{A}_{ps}} \neq \emptyset$ then $\mathbf{0}^q_{\mathcal{A}_{ps}} = \emptyset$. If $\mathbf{1}^d_{\mathcal{A}_{ps}} \neq \emptyset$ this implies that there is at least one $S \in \mathcal{S}$ such that for all $T \in \mathcal{S}$ $\vdash_{\mathcal{A}_{ps}} S \Box\!\!\rightsquigarrow T$. Consequently, there is no $T \in \mathcal{S}$ that no situation S is about, which implies that $\mathbf{0}^q_{\mathcal{A}_{ps}} = \emptyset$. This proves the proposition.

□

Structure of this chapter

In the following sections we present different IR models. Every individual model is described in four subparts entitled: (1) the model, (2) translation, (3) postulates and (4) reflection.

Each time we start a presentation of an IR model in the section ‘The model’. Here we formulate the statement “d about q if and only if ...” in terms of the model. Next, in the section ‘Translation’ we present the translation of the model to our framework. In the section ‘Postulates’, axioms and rules are distilled from the properties of the given IR model. With respect to the model, we inspect the soundness and completeness of the aboutness proof system.

In the last section of each model entitled ‘Reflection’, we elaborate on the following two aspects:

(i) What are the bottom and top elements of the aboutness proof system and the IR model?

(ii) We suggest some improvements of the IR model under scrutiny. Usually an improvement can be obtained by adding or modifying axioms and rules. These new/changed postulates are no longer based on the given IR model, but are defined in terms of the framework.

4.1 Strict coordinate retrieval

The model

The first model we analyse is the so-called *strict coordinate retrieval model.* The matching function that drives strict coordinate retrieval determines the existence of a subset, given the set of descriptors representing a document d and the set of descriptors comprising the query q. The way of interpreting aboutness is by declaring that d is about q if and only if the descriptors of q are a subset of the descriptors of the representation of document d, that is, of $\chi(d)$.

Definition 4.3 Let $\mathcal{D}$ be a document-base and d a document with $d \in \mathcal{D}$. Furthermore, suppose that $\mathcal{T}$ is some finite set of basic information items (descriptors) such that $\chi(d)$ and q are subsets of $\mathcal{T}$, where $\chi(d)$ represents the descriptor set of document d and q is a query. The *strict coordinate aboutness decision* is defined as follows:

$$\models_{\mathrm{SC}_m} d \text{ about } q \quad \text{if and only if } \chi(d) \supseteq q.$$

Translation

In order to translate strict coordinate retrieval to the framework a *basic infon language* $\mathcal{I}_{Basic}(\mathcal{T})$ as given in Definition 3.4 is used. The input of this language is a set of unspecified basic informations items $\mathcal{T}$ that can be almost anything, for instance, keywords, noun-phrases, photo’s, etc.

Definition 4.4 The *basic infon language* $\mathcal{I}_{Basic}(\mathcal{T})$ is the infon language $\mathcal{I}(\mathcal{P}^{+}(\mathcal{T}), \emptyset, \emptyset)$, or alternatively, the language $\mathcal{P}^{+}(\mathcal{T})$.

Such a language contains a set of positive profons based on a set of basic information items. The language of situations $\mathcal{S}_{Basic}$ is the language $\mathcal{S}(\mathcal{I}_{Basic}(\mathcal{T}))$.

The translation of a document representation of a given document-base $\mathcal{D}$ in a situation of $\mathcal{S}_{Basic}$ is defined as follows:

$$map(\chi(d)) = \{\langle\langle \mathrm{I,t;1} \rangle\rangle \mid t \in \chi(d)\}.$$

The translation of a query to a query situation is defined in a similar way. Trivially the function map is bijective modulo set-equivalence. This can be proved by induction, observing that every set with a unique basic information item corresponds to a singleton set containing a profon and vice versa.

Postulates

Next, we propose the underlying aboutness proof system of strict coordinate retrieval, denoted by SC_{ps}.

Definition 4.5 (Strict Coordinate Situation Aboutness) The aboutness proof system SC_{ps} is defined to be the triple $\langle \mathcal{L}(\mathcal{I}_{Basic}(\mathcal{T}))$, {Reflexivity},{Set Equivalence,Left Monotonic Union,Cut}$\rangle$.

The axiom and rules are as given in Chapter 3. The rule Set Equivalence uses the set equivalence relation $\equiv$, which is up till now undefined. This rule expresses the requirement that equivalent sets behave identically with respect to aboutness decisions. Here, the set equivalence relation $\equiv$ is defined as follows:

$$S \equiv T =^{def} (\phi \in S \Leftrightarrow \phi \in T) \text{ for all } \phi \in \mathcal{I}_{Basic}(\mathcal{T}) \text{ and } S, T \in \mathcal{S}_{Basic}.$$

As one may have noticed, we can introduce an aboutness proof system consisting of a single rule, namely the rule Subset Aboutness defined by:

Subset Aboutness (SA)

$$\frac{S \equiv T \cup U}{S \,\Box\!\!\rightsquigarrow T}$$

Here, we want to remind the reader that our intention is not to compress postulates into the smallest possible set, but that our interest is focused on an exploration of some general basic assumptions of aboutness. Therefore, we feel that it is better to present a larger set of axioms and rules to describe the intuition of the model's aboutness decision than to present a small set. Of course, we require that the proof system does not contain derivable axioms or rules, nor should it contain axioms or rules which do not contribute to the understanding of the derivation of aboutness. As one may see, the rule Subset Aboutness does not contribute to the understanding of aboutness.

Theorem 4.1 The aboutness proof system SC_{ps} is sound. That is, for all subsets A, B of $\mathcal{T}$ and $D \in \mathcal{D}$ such that $\chi(D) = A$: if $\vdash_{\mathrm{SC}_{ps}} map(A) \Box\!\!\rightsquigarrow map(B)$ then $\models_{\mathrm{SC}_m} D$ about B.

Proof First we show that the axiom Reflexivity is sound. Secondly we show that all rules are sound. This enables us to conclude that SC_{ps} is sound with respect to SC_m.

- The soundness of the axiom Reflexivity: $S \Box\!\!\rightsquigarrow S$ can be proved as follows. By the definition of the function map we have that if $S \equiv map(A)$ and $S \equiv map(B)$ then $A \equiv B$. If A is equivalent to B, then the aboutness decision that $A \supseteq B$ is sound. This proves the soundness of the axiom Reflexivity.
- Note that Set Equivalence consists of two rules. Here, we show only the soundness of the rule Left Set Equivalence; the proof of soundness of the rule Right Set Equivalence proceeds analogously. Assume that the premises of the rule are sound, that is, $map(A) \equiv map(B)$ and $map(A) \Box\!\!\rightsquigarrow map(C)$ are valid. Then by the definition of the function map we have that $A \equiv B$. Furthermore, the sound assumption $map(A) \Box\!\!\rightsquigarrow map(C)$ allows us to conclude that $A \supseteq C$. We have to inspect whether the conclusion of the Set Equivalence rule $map(B) \Box\!\!\rightsquigarrow map(C)$ is sound. Trivially, if $A \equiv B$ and $A \supseteq C$, then $B \supseteq C$, which implies that $map(B) \Box\!\!\rightsquigarrow map(C)$. This proves the soundness of the Set Equivalence rule.
- In order to prove that Left Monotonic Union is sound, one has to prove that given that $S \Box\!\!\rightsquigarrow T$ is sound, $S \cup U \Box\!\!\rightsquigarrow T$ is a sound conclusion. Assume that $S \equiv map(A)$, $T \equiv map(B)$ and $S \cup U \equiv map(C)$. The sound premise implies that $A \supseteq B$. By the definition of the function map we have that $C \supseteq A$. So, from the fact that $A \supseteq B$ and $C \supseteq A$, the conclusion that $C \supseteq B$ follows directly, which implies that $map(C) \Box\!\!\rightsquigarrow map(B)$. This proves the soundness of the rule Left Monotonic Union.
- Finally, we have to prove the soundness of the Cut rule. Assume that $S \cup T \Box\!\!\rightsquigarrow U$ and $S \Box\!\!\rightsquigarrow T$ are sound premises. Let $S \equiv map(A)$, $T \equiv map(B)$, and $U \equiv map(C)$. Given this premise, we have to prove that $S \Box\!\!\rightsquigarrow U$ is a sound conclusion, that is, $A \supseteq C$. Therefore, we have to inspect whether if $A \cup B \supseteq C$ and $A \supseteq B$ then $A \supseteq C$ is valid. This is obviously true, which proves the soundness of the Cut rule.

□

Theorem 4.2 The aboutness proof system SC_{ps} is complete. That is, for all subsets A, B of $\mathcal{T}$ and $D \in \mathcal{D}$ such that $\chi(D) = A$: if $\models_{\mathrm{SC}_m} D$ about B then $\vdash_{\mathrm{SC}_{ps}} map(A) \Box\!\!\rightsquigarrow map(B)$.

Proof We have to show that if $A \supseteq B$ then $map(A) \Box\!\!\rightsquigarrow map(B)$. Assume $A \supseteq B$, let C be defined by $A \setminus B$. Furthermore, let $map(A) = S$, $map(B) = T$ and $map(C) = U$. By the definition of the function map, $S \equiv T \cup U$. Starting from the Reflexivity axiom one can now make the following derivation:

$$\dfrac{\dfrac{T \Box\!\!\rightsquigarrow T}{T \cup U \Box\!\!\rightsquigarrow T}\,\textsf{LMU} \qquad S \equiv T \cup U}{S \Box\!\!\rightsquigarrow T}\,\textsf{SE}$$

□

The reader may have noticed that the Cut rule is not used in the completeness proof and therefore could be omitted as a postulate for aboutness proof system SC_{ps}. However, we already showed that the Cut rule is sound with respect to the strict coordinate model so adding the rule to the sound and complete system containing the postulates Reflexivity, Set Equivalence, and Left Monotonic Union will not make the proof system overcomplete. The reader may verify that Cut is not derivable from the set of postulates. This type of rule is known as an *admissible* rule. The Cut rule is very useful for aboutness proofs and describes also the intuition of the model's aboutness decision. For this reason we have added the rule to the aboutness proof system SC_{ps}.

Reflection

The axiomatisation of the strict coordinate model gives us the possibility to determine the top and bottom elements as described in Definition 4.1 on page 64.

Since we have proved that SC_{ps} is a sound and complete system for IR model SC_m, we only present the top and bottom elements of SC_{ps}. Because, using Proposition 4.1 the top and bottom elements of SC_m can be derived from the top and bottom elements of SC_{ps}.

Proposition 4.3 In the aboutness proof system SC_{ps} we have that:

(i) The top query of SC_{ps} is the set $\{\emptyset\}$.

(ii) The top document of SC_{ps} is the set $\{\mathcal{I}_{Basic}(\mathcal{T})\}$.

(iii) The bottom query of SC_{ps} is the set $\emptyset$.

(iv) The bottom document of SC_{ps} is the set $\emptyset$.

Proof

(i) We first show that, for every arbitrary situation $S \in \mathcal{S}_{Basic}$, the aboutness formula $S \Box\!\!\rightsquigarrow \emptyset$ is an aboutness theorem. To see this, start from the Reflexivity axiom and observe that one can make the following derivation:

$$\dfrac{\dfrac{\emptyset \Box\!\!\rightsquigarrow \emptyset}{\emptyset \cup S \Box\!\!\rightsquigarrow \emptyset}\,\mathsf{LMU} \qquad \emptyset \cup S \equiv S}{S \Box\!\!\rightsquigarrow \emptyset}\,\mathsf{SE}$$

Furthermore, we have to show that there is, beside the empty set, no other element T in $\mathbf{1}^{q}_{\mathrm{SC}_{ps}}$ such that for all situations S, $S \Box\!\!\rightsquigarrow T$. Assume there is. Since $T \neq \emptyset$, it should contain at least one element. So, $T \equiv \{\varphi\} \cup U$ with U possibly empty. Consider a situation S with $\varphi \notin S$. Using the completeness theorem one can see that it is not possible to deduce that $S \Box\!\!\rightsquigarrow T$, which is a contradiction.

(ii) We show that the situation $\mathcal{I}_{Basic}(\mathcal{T})$ is about all situations $T \in \mathcal{S}_{Basic}$. First we note that $\mathcal{I}_{Basic}(\mathcal{T})$ is here viewed as a situation that contains all the infons of the language $\mathcal{I}_{Basic}(\mathcal{T})$. So, the situation $\mathcal{I}_{Basic}(\mathcal{T})$ is the situation with the maximal number of elements. For every $T \in \mathcal{S}_{Basic}$ it is provable from $T \Box\!\!\rightsquigarrow T$, that $T \cup \mathcal{I}_{Basic}(\mathcal{T}) \Box\!\!\rightsquigarrow T$. Using Set Equivalence one determines that $\mathcal{I}_{Basic}(\mathcal{T}) \Box\!\!\rightsquigarrow T$. So, indeed, $\mathcal{I}_{Basic}(\mathcal{T})$ is about every situation of $\mathcal{S}_{Basic}$. Furthermore, we have to show that there is, beside the situation $\mathcal{I}_{Basic}(\mathcal{T})$, no other element meeting this requirement. If this were the case, at least one infon φ should not be in the situation. Assume such a situation S. This situation S can not be about the situation $\{\varphi\}$ which is a contradiction.

(iii) Since $\mathbf{1}^{d}_{\mathrm{SC}_{ps}}$ is not empty, Proposition 4.2 suffices to conclude that $\mathbf{0}^{q}_{\mathrm{SC}_{ps}}$ is empty.

(iv) Since $\mathbf{1}^{q}_{\mathrm{SC}_{ps}}$ is not empty, Proposition 4.2 suffices to conclude that $\mathbf{0}^{d}_{\mathrm{SC}_{ps}}$ is empty.

□

In terms of queries and documents: if somebody enters an empty query all the documents are retrieved[1]. The document that is indexed with all descriptors of the descriptor set will always be retrieved, as $map(\mathcal{T}) = \mathcal{I}_{Basic}(\mathcal{T})$. For each query there is always a relevant document representation[2]. Conversely, for each document, it is always possible to construct a query that will retrieve the document.

We see several directions in which the aboutness proof system SC_{ps} could be extended. First of all, additional knowledge can be adopted in the system as axioms of the type $\varphi \rightarrow \psi$, in addition with the Union Containment rule. Then one has the ability to express informational containment relations between basic information items, in order to retrieve more relevant documents.

[1] Note also that in this theoretical case, we assume that it is possible to index a document with an empty set. In this particular case, there are no descriptors available that present the contents of the document correctly.

[2] This does not imply that there is always a relevant document, since the representation set $\mathcal{T}$ may not correspond to a document.

4.2 Coordinate retrieval

As the term 'strict' in strict coordinate retrieval implies, the aboutness decisions of SC_{ps} are 'strict', in the sense that there are only a few possibilities to derive aboutness. As a result, only a few documents are considered to be relevant in a strict coordinate retrieval system. In order to deliver some more documents, which still are very likely to be relevant, one could adopt a rule that allows us to extend the right-hand side of the aboutness relation. This is the case in the following model we consider, the *coordinate retrieval* model.

The model

The matching function which drives coordinate retrieval determines overlap. A document d is about q if and only if there is some overlap between the representations of d and q.

Definition 4.6 Let $\mathcal{D}$ be a document-base and d a document with $d \in \mathcal{D}$. Furthermore, let $\mathcal{T}$ be some finite set of basic information items (descriptors) such that $\chi(d)$ and q are subsets of $\mathcal{T}$, where $\chi(d)$ represents the descriptor set of document d and q a query. The *coordinate aboutness decision* is defined as follows:

$$\models_{\mathrm{C}_m} d \text{ about } q \quad \text{if and only if } \chi(d) \cap q \not\equiv \emptyset.$$

The aboutness relation is *symmetric*: if there is an overlap between S and T then obviously there is an overlap between T and S.

Translation

The mapping of coordinate retrieval representatives to our framework proceeds in the same way as for strict coordinate retrieval. Here, the *basic infon language* $\mathcal{I}_{Basic}(\mathcal{T})$ as given in Definition 4.4 is used. Both document representations and queries are modelled as situations that are elements of the language $\mathcal{S}_{Basic}$. The equivalence relation is defined analogously to the one of aboutness proof system SC_{ps}.

Postulates

The aboutness proof system for aboutness decisions in coordinate retrieval is denoted by C_{ps} and is defined as follows.

Definition 4.7 (Coordinate Situation Aboutness) The aboutness proof system C_{ps} is defined to be the triple $\langle \mathcal{L}(\mathcal{I}_{Basic}(\mathcal{T})), \{\text{Singleton Reflexivity}\}, \{\text{Set Equivalence}, \text{Left Monotonic Union}, \text{Symmetry}, \text{Strict Composition}\}\rangle$.

The attentive reader may have noticed that we did not adopt the Reflexivity axiom. We did not do this for the following reason: it would enable us to prove that S is about T for arbitrary situations S and T. To see this, observe the following. First, Reflexivity holds for empty sets, that is, $\emptyset \Box\!\!\rightsquigarrow \emptyset$. Given Left Monotonic Union, Symmetry and Set Equivalence, it is then provable that $S \Box\!\!\rightsquigarrow T$ for arbitrary situations S and T, as showed by the following prooftree:

$$\dfrac{\dfrac{\dfrac{\dfrac{\dfrac{\emptyset \Box\!\!\rightsquigarrow \emptyset}{\emptyset \cup T \Box\!\!\rightsquigarrow \emptyset}\,\mathsf{LMU}}{\emptyset \Box\!\!\rightsquigarrow \emptyset \cup T}\,\mathsf{Sy}}{\emptyset \cup S \Box\!\!\rightsquigarrow \emptyset \cup T}\,\mathsf{LMU} \qquad \emptyset \cup S \equiv S}{S \Box\!\!\rightsquigarrow \emptyset \cup T}\,\mathsf{SE} \qquad \emptyset \cup T \equiv T}{S \Box\!\!\rightsquigarrow T}\,\mathsf{SE}$$

In order to avoid this kind of anomaly we adopt a special version of the Reflexivity axiom called Singleton Reflexivity as presented in Chapter 3.

Note that C_{ps} covers the Strict Composition rule instead of the Composition rule of the SC_{ps} aboutness proof system. With the Composition rule it was allowed, given the assumption $S \Box\!\!\rightsquigarrow T$, to reduce the right-hand side of the aboutness relation taking the intersection $T \cap U$, where U is an arbitrary situation. With Strict Composition one can only reduce the right-hand side by taking the intersection $T \cap S$. The premise of this rule is necessary in order to avoid that $S \cap T$ could result in an empty set, which would lead to the same problem as with Reflexivity.

Theorem 4.3 The aboutness proof system C_{ps} is sound. That is, for all subsets A, B of $\mathcal{T}$ and $D \in \mathcal{D}$ such that $\chi(D) = A$: if $\vdash_{\mathrm{C}_{ps}} map(A) \Box\!\!\rightsquigarrow map(B)$ then $\models_{\mathrm{C}_m} D$ about B.

Proof Firstly we prove the soundness of the axiom Singleton Reflexivity. Secondly we prove the soundness of all the rules of C_{ps}. This enable us to conclude that C_{ps} is sound with respect to C_m.

- The axiom Singleton Reflexivity is sound. We have to show that, if $map(A) \equiv \{\varphi\}$ and $map(B) = \{\varphi\}$, then $A \cap B \not\equiv \emptyset$. By the definition of the function map we have that given $map(A) \equiv map(B) \equiv \{\varphi\}$ for some $\varphi \in \mathcal{I}_{Basic}(\mathcal{T})$, then $A \equiv B \equiv \{t\}$ for some $t \in \mathcal{T}$. Consequently, $A \cap B \equiv \{t\}$, which proves the soundness of the axiom.
- The Set Equivalence rule is sound. Similar as with the soundness proof of SC_{ps}, we show only the soundness of the rule Left Set Equivalence. Given that $map(A) \equiv map(B)$ and $map(A) \Box\!\!\rightsquigarrow map(C)$ are sound premises, which implies that $A \equiv B$ and $A \cap C \not\equiv \emptyset$, we have to inspect whether the conclusion $map(B) \Box\!\!\rightsquigarrow map(C)$ is sound. Trivially, if $A \equiv B$ and $A \cap C \not\equiv \emptyset$, then $B \cap C \not\equiv \emptyset$. This proves the soundness of the Set Equivalence rule.

- In order to prove that Left Monotonic Union is sound, one has to prove that given that $S \Box\!\!\!\rightsquigarrow T$ is a sound premise, the conclusion $S \cup U \Box\!\!\!\rightsquigarrow T$ is sound. Let $S \equiv map(A)$, $T \equiv map(B)$ and $S \cup U \equiv map(C)$. The sound premise $S \Box\!\!\!\rightsquigarrow T$ implies that $A \cap B \not\equiv \emptyset$. By the definition of the function map we have that $C \supseteq A$. So, the fact that $C \supseteq A$ and $A \cap B \not\equiv \emptyset$, the conclusion that $C \cap B \not\equiv \emptyset$ follows directly. This proves the soundness of the rule Left Monotonic Union.
- Finally, we prove soundness of Strict Composition, that is, given that $S \Box\!\!\!\rightsquigarrow T$ is a sound premise, $S \Box\!\!\!\rightsquigarrow T \cap S$ should be a sound conclusion. Let $S \equiv map(A)$ and $T \equiv map(B)$. Given the premise that $A \cap B \not\equiv \emptyset$ the conclusion that $A \cap B \cap A \not\equiv \emptyset$ is valid. Set-theoretically we have that $A \cap B \equiv (A \cap B) \cap A$, so the conclusion $S \Box\!\!\!\rightsquigarrow T \cap S$ is valid, and consequently the rule Strict Composition is sound.

□

Similar to the proof of Theorem 4.1, we remark that it is possible to omit the rule Strict Composition as a postulate for aboutness proof system C_{ps}. We have added the admissible rule for the same reasons as we gave for the Cut rule of the aboutness proof system SC_{ps}.

Theorem 4.4 The aboutness proof system C_{ps} is complete. That is, for all subsets A, B of $\mathcal{T}$ and $D \in \mathcal{D}$ such that $\chi(D) = A$: if $\models_{\mathrm{C}_m} D$ about B then $\vdash_{\mathrm{C}_{ps}} map(A) \Box\!\!\!\rightsquigarrow map(B)$.

Proof We have to show that if $A \cap B \not\equiv \emptyset$ then $map(A) \Box\!\!\!\rightsquigarrow map(B)$. Assume $A \cap B \not\equiv \emptyset$. Then obviously a singleton set C exists such that $C \subseteq A$ and $C \subseteq B$. Let $D \equiv A \setminus C$ and $E \equiv B \setminus C$. Furthermore, let $map(A) = S$, $map(B) = T$, $map(C) = U$, $map(D) = V$, and $map(E) = W$. Then consequently, $map(A) = map(C) \cup map(D)$ and so on. Starting with Singleton Reflexivity, we find:

$$\dfrac{\dfrac{\dfrac{\dfrac{\dfrac{U \Box\!\!\!\rightsquigarrow U}{U \cup W \Box\!\!\!\rightsquigarrow U}\,\mathsf{LMU} \qquad T \equiv U \cup W}{T \Box\!\!\!\rightsquigarrow U}\,\mathsf{SE}}{U \Box\!\!\!\rightsquigarrow T}\,\mathsf{Sy}}{U \cup V \Box\!\!\!\rightsquigarrow T}\,\mathsf{LMU} \qquad S \equiv U \cup V}{S \Box\!\!\!\rightsquigarrow T}\,\mathsf{SE}$$

□

Reflection

As our framework is developed with the intention to compare models, we will inspect now what the top and bottom elements of the aboutness proof system C_{ps} are and detect the differences with those of the aboutness proof system SC_{ps}.

Proposition 4.4 In the aboutness proof system C_{ps} we have that:

(i) The bottom query of C_{ps} is the set $\{\emptyset\}$.

(ii) The bottom document of C_{ps} is the set $\{\emptyset\}$.

(iii) The top query of C_{ps} is the set $\emptyset$.

(iv) The top document of C_{ps} is the set $\emptyset$.

Proof To give the proof the following claim is needed:

Claim 4.1 For all situations $S \in \mathcal{S}_{Basic}$, $\not\vdash_{\mathrm{C}_{ps}} S \,\Box\!\!\rightsquigarrow \emptyset$.

Proof Using the soundness theorem 4.3 and completeness theorem 4.4, we can infer that $\vdash_{\mathrm{C}_{ps}} map(A) \,\Box\!\!\rightsquigarrow map(B)$ if and only if $A \cap B \not\equiv \emptyset$. Hence it follows that $\not\vdash_{\mathrm{C}_{ps}} map(A) \,\Box\!\!\rightsquigarrow map(B)$ if and only if $A \cap B \equiv \emptyset$. By definition $map(\emptyset) = \emptyset$ and for all $S \in \mathcal{S}_{Basic}$, $S \cap \emptyset \equiv \emptyset$. This suffices to conclude that for all situations $S \in \mathcal{S}_{Basic}$, $\not\vdash_{\mathrm{C}_{ps}} S \,\Box\!\!\rightsquigarrow \emptyset$.

□

To complete the proof of Proposition 4.4, we use Claim 4.1. Since cases (ii) and (iv) are, due to the symmetry property, analogous to cases (i) and (iii) respectively, we restrict ourselves to proving the items (i) and (iii).

(i) The fact that $\emptyset \in \mathbf{0}^{q}_{\mathrm{C}_{ps}}$ is shown by the claim. Furthermore, we have to show that besides the empty-set, there is no other element S meeting the requirement of the elements of the bottom query set. Assume there is. Since $S \not\equiv \emptyset$, $S \equiv \{\varphi\} \cup U$ for some φ and U. If U is $\emptyset$, then S is about $\{\varphi\}$ by Singleton Reflexivity otherwise we can use Left Monotonic Union to conclude, starting from $\{\varphi\} \,\Box\!\!\rightsquigarrow \{\varphi\}$, that $S \,\Box\!\!\rightsquigarrow \{\varphi\}$. Hence there is no other situation for which no situation is about.

(iii) Since $\mathbf{0}^{d}_{\mathrm{C}_{ps}}$ is not empty, Proposition 4.2 suffices to conclude that $\mathbf{1}^{q}_{\mathrm{C}_{ps}}$ is empty.

□

In terms of queries and documents this implies that for coordinate retrieval, in contrast with strict coordinate retrieval, an empty query will never retrieve any document, not even those documents that are represented by an empty set. These kind of documents are never retrieved. Consequently, it is not possible to construct a query that will retrieve all documents, or to index a document in such a way that it will always be retrieved.

The aboutness proof system C_{ps} can also be extended by adopting knowledge axioms of the type $\varphi \rightarrow \psi$ in addition to the Union Containment. Maybe, with this system there are too many possibilities to derive aboutness. As a result, too many documents are considered to be relevant in a coordinate retrieval system. In order to deliver fewer documents, we could adopt some guarded rules as presented in Chapter 3, instead of Left Monotonic Union.

4.3 Vector-Space retrieval

The model

The vector-space model originates from the work of Salton [131]. As the name of the model indicates, vector-space retrieval adopts a geometric viewpoint. The set of descriptors $\mathcal{T}$ is ordered (mostly alphabetically). The list of descriptors is then used to represent a n-dimensional space, where n is the total of number of descriptors in $\mathcal{T}$. The descriptor set of a document or a query is transformed to a vector as follows.

Let $\mathcal{T}$ be a finite descriptor set with n descriptors $k_1, \ldots k_n$ (in this fixed order). For document d the vector is $\langle t_1, \ldots, t_n \rangle$ with $t_i = 1$ if $k_i \in \chi(d)$ and 0 otherwise. Similarly for the query q one can construct the vector $\langle u_1, \ldots, u_n \rangle$. Note, that in the mapping process from descriptor sets onto vectors, the order of the elements in the descriptor sets does not play a role. The only requirement is that the set $\mathcal{T}$ should be represented in a fixed order.

In Salton's model the relevance of a document d given a query q is estimated using the cosine of the angle between the two vectors of d and q.

Let t be the vector of $\chi(d)$ and u the vector of q, then the estimation of relevance is based on the following relevance cosine-function:

$$relcos(\chi(d), q) = \frac{t \cdot u}{\parallel t \parallel \cdot \parallel u \parallel} = \frac{\sum_{i=1}^{n} t_i \times u_i}{\sqrt{\sum_{i=1}^{n} t_i^2} \times \sqrt{\sum_{i=1}^{n} u_i^2}}.$$

In order to avoid undefined cases, we define $relcos(\chi(d), q)$ to be 0 if $\chi(d)$ or $q = \emptyset$.

Rather than using binary values for the vectors, term weights can be used as descriptive values for the descriptors. Typically these term weights are based on occurrence frequencies [1]. The term weight is a value between 0 and 1. The estimation of relevance remains the same. However, in this section only binary values are used.

Definition 4.8 Let $\mathcal{D}$ be a document-base and d be a document with $d \in \mathcal{D}$. Furthermore, suppose that $\chi(d)$ and q are subsets of $\mathcal{T}$, where $\chi(d)$ represents the descriptor set of document d and q a query. The ***vector-space aboutness decision*** is defined as follows:

$$\models_{\mathrm{VC}_m} d \text{ about } q \quad \text{if and only if } relcos(\chi(d), q) > 0.$$

Here, we have fixed the aboutness definition of a vector-space model in terms of the cosine being greater than zero. Another suggestion could be that d is about q if and only if $relcos(\chi(d), q) = 1$. This would imply an IR system that returns very few relevant documents. Or, we could use a 'cut-off'-value based on the following lemma.

Lemma 4.1 Let $\mid A \cap B \mid = k$. Then

$$relcos(A, B) \geq \frac{k}{\frac{1}{2}(n + k)}.$$

Proof In the function *relcos*, A and B are mapped onto $0 \Leftrightarrow 1$ vectors that coincide in k ones. Let A have n_a more 1's and B have n_b more 1's, with clearly $n_a + n_b \leq n \Leftrightarrow k$. Then

$$\begin{aligned} relcos(A,B) &= \frac{k}{\sqrt{(k+n_a)(k+n_b)}} = \frac{k}{\sqrt{k^2 + k(n_a+n_b) + n_a n_b}} \\ &\geq \frac{k}{\sqrt{k^2 + k(n \Leftrightarrow k) + n_a n_b}} = \frac{k}{\sqrt{kn + n_a n_b}} \\ &\geq \frac{k}{\sqrt{kn + \frac{1}{4}(n \Leftrightarrow k)^2}} = \frac{k}{\sqrt{\frac{1}{4}(n+k)^2}} = \frac{k}{\frac{1}{2}(n+k)}. \end{aligned}$$

□

This bound is pretty good because there always are A and B that match it, up to small rounding errors. In Chapter 5 we will look at alternatives for modelling different aboutness-levels (for instance, aboutness$_1$ with $relcos(\chi(d), q) = 1$ and aboutness$_2$ with $relcos(\chi(d), q) > 0$) without changing the structure of the underlying aboutness proof system. But before we pursue this, we continue the discussion of the vector-space model in our framework.

Translation

The mapping of vector-space retrieval representatives to our framework proceeds in a way similar to strict coordinate retrieval and coordinate retrieval. Again, the basic infon language $\mathcal{I}_{Basic}(\mathcal{T})$ as defined in Definition 3.4 is used. Both document representations and queries are modelled as situations that are elements of the language $\mathcal{S}_{Basic}$. The equivalence relation is defined analogously to the one of aboutness proof system SC$_{ps}$.

Postulates

For the aboutness derivations in the vector-space model we propose the following aboutness proof system:

Definition 4.9 (Vector-Space Situation Aboutness) The aboutness proof system VC$_{ps}$ is defined to be the triple $\langle \mathcal{L}(\mathcal{I}_{Basic}(\mathcal{T}))$, {Singleton Reflexivity},{Set Equivalence,Left Monotonic Union,Symmetry,Strict Composition}$\rangle$.

This aboutness proof system is identical to aboutness proof system C$_{ps}$. The soundness and completeness theorems of VC$_{ps}$ can be presented in different ways. One way of presenting the theorems is based on the following lemma:

Lemma 4.2 For a given set of descriptors $A, B \subset \mathcal{T}$,

$$relcos(A,B) > 0 \Leftrightarrow A \cap B \not\equiv \emptyset.$$

Proof

$\Rightarrow$ Assume that $relcos(A, B) > 0$. Because vectors have nonnegative coordinates this implies that $\sum_{i=1}^{n} t_i \times u_i > 0$ with t the vector of A and u the vector of B and thus there exists an $i \leq n$ such that $t_i \times u_i \neq 0$. The result of $t_i \times u_i$ is zero if the descriptor k_i occurs in A and not in B or vice versa. Thus, if $t_i \times u_i \neq 0$ there is a descriptor k_i that is both an element of A and of B. To conclude, if $relcos(A, B) > 0$ then A and B are having one or more descriptors in common, and thus $A \cap B \not\equiv \emptyset$.

$\Leftarrow$ Assume that $A \cap B \not\equiv \emptyset$, thus $\mid A \cap B \mid = k$ for $k > 0$. The fact that $relcos(A, B) > 0$ is shown by Lemma 4.1. Since $n, k > 0$, we have that $\frac{k}{\frac{1}{2}(n+k)} > 0$. This suffices to conclude that $relcos(A, B) > 0$.

□

The lemma shows that the aboutness decision of the vector-space model and the coordinate retrieval model are equivalent. The soundness and the completeness of VC_{ps} for the model VC_m then follows directly.

Corollary 4.2

(i) The aboutness proof system VC_{ps} is sound. That is, for all subsets A, B of $\mathcal{T}$: if $\vdash_{\mathrm{VC}_{ps}} map(A) \Box\!\!\rightsquigarrow map(B)$ then $\models_{\mathrm{VC}_m} A$ about B.

(ii) The aboutness proof system VC_{ps} is complete. That is, for all subsets A, B of $\mathcal{T}$: if $\models_{\mathrm{VC}_m} A$ about B then $\vdash_{\mathrm{VC}_{ps}} map(A) \Box\!\!\rightsquigarrow map(B)$.

Reflection

As VC_{ps} is identical to C_{ps}, the top and bottom elements of VC_{ps} are identical to those of the aboutness proof system C_{ps}.

4.4 Index Expression Belief Network retrieval

The use of probabilistic laws for information retrieval to determine whether a document is about a query is considered to be both elegant and potentially extremely powerful [122]. The main argument for adopting a probabilistic definition for relevance is that for the determination of relevance one must use imperfect knowledge; the query is not an exact match with the information need and the document representation is only a crude approximation of the document content. Aboutness derivations are in this view embodied by a probabilistic reasoning process. Typically, in a probabilistic retrieval model one decides that d is about q if and only if the estimation of the probability of a document d given a query q is larger than a cut-off value x, denoted as $P(\chi(d) \mid q) > x$

with P the probability function. Various probabilistic information retrieval models have been proposed. One particular class consists of the network-based probabilistic retrieval models [23, 24, 31, 32, 58, 77, 143].

The model

In this section we describe one particular network-based probabilistic retrieval model, namely the *Index Expression Belief Networks (IEBN)* [23, 24, 77]. As the name of the model indicates, an IEBN-model contains two aspects: *Index Expressions* and *Belief Networks*.

First we briefly introduce the notion of belief networks. A belief network is a graphical representation of a problem domain depicting the probabilistic variables of the domain and their interdependencies. Belief networks are used to calculate the belief in the occurrence of an event[3]. For instance, a belief network can be used as a diagnostic system in order to quantify the belief that someone has fever given the fact that someone has a high temperature.

A belief network is a directed acyclic graph with a set of nodes V_G consisting of probabilistic variables representing the belief in an event, and a set of edges E_G each representing the interdependency between two events. Furthermore for each node n of the graph there are *assessment functions* $\gamma_n(x)$ that represent the initial belief in each event n to be true (denoted as x) or false (denoted as $\neg x$). For instance, the belief that someone has fever could be set to 0.15 (denoted as $\gamma_{\text{Fever}}(\mathit{Fever}) = 0.15$).

In case we know that a person has a high temperature, then we could assume that it is more likely that the person has fever. All the conditional factors (such as red colour, sweating and so on) are set by the assessment functions. The complete set of all assessment functions is denoted by Γ.

The calculation of the belief in a certain event proceeds as follows. We enter some evidence in the network, that is, things we know for sure. For example, we know that the person has a high temperature and a red colour. We recalculate the nodes based on the probability distribution expressed by the directed graph (Pearl [115] designed various algorithms to perform this calculation efficiently) and then 'read' the belief-factor of a node. This calculation process is called *evidence propagation*.

In the IEBN-model, the belief network approach is used to calculate our belief that a document is about a query. Here, the graphical representation of the belief network is based on the index expressions. Each document d in the document-base is represented as a set of index expressions $\chi(d)$. From the union of all representation sets a directed graph, termed a *lattice* [23], is constructed. This lattice is used as the graph of the belief

[3] See e.g. Pearl's book [115] for a detailed presentation of belief networks.

network. Next, we show how we can construct a lattice out of the index expressions, that can be used to calculate the probability that d is about q.

In Chapter 3 at page 41, we already introduced index expressions as elements of a relational indexing approach. Here, we introduce the definition of index expressions based on the definition of Bruza [23]:

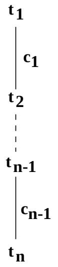
Definition 4.10 Let $\mathcal{T}$ be a set of descriptors and C a set of connectors. The language $\mathcal{L}(\mathcal{T}, C)$ of index expressions is defined by:

- for $n \in \mathbb{N}$ with $0 \leq i \leq n$ and $t_i \in \mathcal{T}$ and $c_i \in C$: $(t_1\ c_1\ t_2 \ldots t_{n-1}\ c_{n-1}\ t_n) \in \mathcal{L}(\mathcal{T}, C)$;
- if $c \in C$ and $I, J \in \mathcal{L}(\mathcal{T}, C)$ then $I\ c\ J \in \mathcal{L}(\mathcal{T}, C)$.

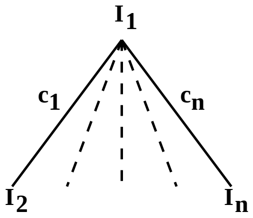
In order to represent an empty index expression (for n=0), the symbol ϵ is included in the language $\mathcal{L}(\mathcal{T}, C)$. Here, the set of descriptors intentionally corresponds to a set of textual elements, for instance, '(cruel ∘ murder of Caesar by Brutus)' is a typical example of an index expression, with the keywords cruel,murder,Caesar,Brutus and the connectors ∘,of,by. Brackets can be used to represent that some connectors bind index expressions stronger than others. Bruza [23] suggested different priorities of the connectors. For instance, the connector ∘ between cruel and murder binds the terms stronger than the connector by between Caesar and Brutus. In this view, the index expression given above could be presented as '(cruel ∘ murder) of (Caesar) by (Brutus)'. Bruza presented an algorithm that based on a priority-list of connectors and given a sentence produces an index expression with brackets included. Brackets will usually be dropped as much as possible without causing confusing.

Bruza presents a function that given an index expression results into a tree-representation of the particular index expression.

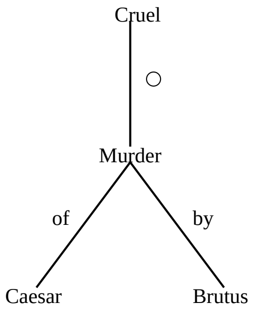
If the index expression I is of the form $(t_1\ c_1\ \ldots c_{n-1}\ t_n)$ then I can be graphically represented as follows:



In case the index expression I is of the form $I_1 c_1 I_2 \ldots c_{n-1} I_n$ with I_i a non-empty index expression for every $1 \leq i \leq n$, the index expression can be depicted as follows:



In this way every index expression can be represented as a tree. For instance the index expression (cruel $\circ$ murder) of Caesar by Brutus can be graphically represented as the following tree:



Index expressions can be ordered based on an 'is-sub-index-expression-of' relation (denoted by $\underline{\subset\!\!\ast}$). The informal definition of this relation, taken from [77], is as follows:

Definition 4.11 Let I_1 and I_2 be index expressions. Then

$I_1 \underline{\subset\!\!\ast} I_2$ if I_1 is a subtree of the tree-representation of I_2.

The set with all the subindex expressions of I is termed the *power index expression* [23], and formally defined as follows:

Definition 4.12 Let I be an index expression in a language $\mathcal{L}(\mathcal{T}, C)$. The *power index expression* of I, denoted by $\wp(I)$, is the set

$$\wp(I) = \{J \mid J \underline{\subset\!\!\ast} I\}$$

where $\underline{\subset\!\!\ast}$ is the is-subexpression-of relation as given in Definition 4.11.

Next we present how a lattice can be constructed out of the power index expressions. In the IEBN-model, every document d of the document-base $\mathcal{D}$ is indexed by a set of index expressions (thus $\chi(d) \subseteq \mathcal{L}(\mathcal{T}, C)$). In the model a directed acyclic graph with a set of nodes V_G of index expressions is constructed using the subindex expression relation to present the edges E_G. The set V_G will be $\bigcup_{d \in \mathcal{D}} \wp(\chi(d))$. The nodes represent the fact that an object represented by index expression is relevant. The edges of a belief network represent an interdependency between two events. Therefore Bruza chooses the relation $\underline{\subset\!\!\ast}$ to be the interdependency relation. The fact that d is about cruel $\circ$ murder depends on conditions such as whether d is about cruel and d is about murder, or not. To capture this intuitive idea the set of edges E_G is defined as follows: $(I_i, I_j) \in E_G$ if and only if $I_i \underline{\subset\!\!\!\ast} I_j$ with $I_i, I_j \in V_G$ and for all $I_k \in V_G$, if $I_i \underline{\subset\!\!\!\ast} I_k$ and $I_k \underline{\subset\!\!\!\ast} I_j$ then $I_k = I_i$ or $I_k = I_j$.

For instance, given the document indexed by one index expression: (cruel $\circ$ murder) of Caesar by Brutus. The following lattice can be constructed:

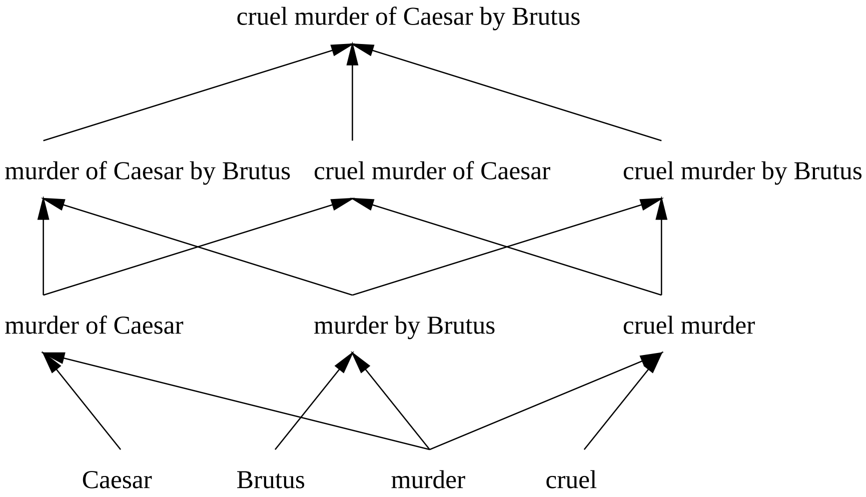


Figure 4.1: A belief network.

This lattice, which is a directed acyclic graph, is used as a belief network, which is called an *IEBN*. Consider the directed graph depicted in Figure 4.1. In a belief network nodes represent events. In the IEBN, the nodes are represented by index expressions. This IEBN captures that the belief that an object is about *'murder by Brutus'* depends on it being about *'murder'* and it being about *'Brutus'*. As mentioned before, besides the directed graph G a belief network consists of a set Γ of probability assessment function, notated as γ. These functions try to assess conditional probabilities. The question is how the nodes of the IEBN could be initiated. For instance, $\gamma_{\text{murder by Brutus}}(\textit{murder by Brutus}) = 0.15$ denotes the assessment function of the node murder by Brutus and intuitively it means that the belief that an object is about *murder by Brutus* is 0.15 (thus the belief that an object is not about *murder by Brutus*, denoted by $\neg$*murder by Brutus* is 0.85). We could also have the following assessment: $\gamma_{\text{murder by Brutus}}(\textit{murder by Brutus} \mid \textit{murder} \wedge \textit{Brutus}) = 0.8$, here we express that the belief in the fact that an object is about murder by Brutus, knowing that an object is about murder and about Brutus is 0.8.

Bruza defines the γ-function of Γ as follows. If a node I has no predecessors, e.g., for all $J \in V_G : (J, I) \notin E_G$ (in the graph these are nodes which are elements of $\mathcal{T}$) then $\gamma_{\mathrm{t}}(t)$ is based on some frequency-value. If a term occurs frequently in a small set of documents then it gets a higher initial value. Typically a frequency function *freq* is used which is a normalised function with its domain between 0 and 1:

$$\gamma_{\mathrm{t}}(t) = \mathit{freq}(t) \text{ and } \gamma_{\mathrm{t}}(\neg t) = 1 \Leftrightarrow \gamma(t).$$

So, the belief that an object is about a keyword depends on the frequency of the keyword in the document-base.

Furthermore Bruza assumed that the probability of a node with predecessors, being true or false, depends on its predecessors. In the *No Blind Faith* theorem [23], Bruza states that the probability is only non-zero if both predecessors are true. Intuitively this is grounded on the fact that an object cannot be about $I_1 c I_2$ if the object is not about I_1 or it is not about I_2.

In case both predecessors are true, the value of this probability assessment depends on which connector is used. For instance, the occurrence of a null-connector $\circ$ between two index expressions is more likely than the connector *around*, i.e., $\gamma_{\mathrm{poor} \,\circ\, \mathrm{Caesar}}(\mathit{poor} \circ \mathit{Caesar} \mid \mathit{poor} \wedge \mathit{Caesar})$ is larger than $\gamma_{\mathrm{poor\ around\ Caesar}}(\mathit{poor\ around\ Caesar} \mid \mathit{poor} \wedge \mathit{Caesar})$. Therefore, Bruza [23] proposed $\gamma_{\varphi \,\circ\, \psi}(\varphi \circ \psi \mid \varphi \wedge \psi) = 0.5366$ and $\gamma_{\varphi \ \mathit{around}\ \psi}(\varphi \ \mathit{around}\ \psi \mid \varphi \wedge \psi) = 0.0017$ based on analysis of the percentage connectors in the underlying document-domain. This analysis can directly be used for the probability estimation. Following [77], we will denote the probability of the occurrence of a connector c as $\mathbf{P}(c)$, so γ is defined by:

$$\gamma_{t_1 \ c \ t_2}(t_1 \ c \ t_2 \mid t_1 \wedge t_2) = \mathbf{P}(c).$$

Furthermore, if two index expressions I and J are not in $\mathcal{T}$, and neither have the form $t_i \ c \ t_j$, and $(I, K), (J, K)$ are elements of E_G then

$$\gamma_{\mathrm{K}}(K \mid I \wedge J) = 1.$$

Using the Shinto-theorem from [77] which states that the a priori probability that a node X has value X fully depends on its 'ancestors', we define the probability function as follows:

Definition 4.13 Let $I \in \mathcal{L}(\mathcal{T}, C)$ and $B = (G, \Gamma)$ be a belief network, and Γ is the set of assessment functions. Furthermore, let $P^+(I)$ be defined by $\wp(I) \setminus \{\epsilon\}$ and $\bigwedge \rho(J) =^{def} K_1 \wedge K_2 \wedge \ldots \wedge K_n$ with K_1 through K_n such that $(K_i, J) \in E_G$. Then:

$$P(I) =^{def} \prod_{J \in P^+(I)} \gamma_{\mathrm{J}}(J \mid \bigwedge \rho(J)).$$

For example, the probability of $P(\text{cruel} \circ \text{murder of Caesar})$, abbreviated as $P(\text{Cr} \circ \text{Mu of Ca})$, is calculated as follows:

$$\begin{aligned} P(\text{Cr} \circ \text{Mu of Ca}) &= \gamma_{\text{Cr} \circ \text{Mu of Ca}}(\text{Cr} \circ \text{Mu of Ca} \mid \text{Cr} \circ \text{Mu} \wedge \text{Mu of Ca}) \times \\ &\quad \gamma_{\text{Cr} \circ \text{Mu}}(\text{Cr} \circ \text{Mu} \mid \text{Cr} \wedge \text{Mu}) \times \\ &\quad \gamma_{\text{Mu of Ca}}(\text{Mu of Ca} \mid \text{Mu} \wedge \text{Ca}) \times \\ &\quad \gamma_{\text{Cr}}(\text{Cr}) \times \gamma_{\text{Mu}}(\text{Mu}) \times \gamma_{\text{Ca}}(\text{Ca}) \\ &= 1 \times \mathbf{P}(\circ) \times \mathbf{P}(\text{of}) \times \mathit{freq}(\text{Cr}) \times \mathit{freq}(\text{Mu}) \times \mathit{freq}(\text{Ca}). \end{aligned}$$

Inspecting this function, we see that the set $\wp(\text{Cr} \circ \text{Mu of Ca})$ bears an 'information bearing index expressions' subset $\{\text{Cr} \circ \text{Mu}, \text{Mu of Ca}, \text{Cr}, \text{Mu}, \text{Ca}\}$. This set contains those elements that influence the probability of the index expression. Formally we define this set as follows:

Definition 4.14 Given an index expression I the *information bearing index expression* subset (denoted by $\wp(I)^+$) is defined to be the smallest subset of I such that

$$\wp(I)^+ =^{def} \{x \mid x \in \mathcal{T} \text{ or } x \text{ has the form } t_1 \; c \; t_2 \text{ with } t_1, t_2 \in \mathcal{T}, c \in C\}.$$

Observing that $P(I \mid I_1 \wedge \ldots \wedge I_k) = 1$ in case I not in $\wp(I)^+$ and $(I_i, I) \in E_G$ for $1 \geq i \geq k$ we propose the following theorem:

Theorem 4.5 Let $I \in \mathcal{L}(\mathcal{T}, C)$ and $B = (G, \Gamma)$ be a belief network, and Γ is the set of assessment functions. Furthermore, $\bigwedge \rho(J) =^{def} K_1 \wedge K_2 \wedge \ldots \wedge K_n$ with K_1 through K_n such that $(K_i, J) \in E_G$. Then:

$$P(I) = \prod_{J \in \wp(I)^+} \gamma_{\text{J}}(J \mid \bigwedge \rho(J)).$$

For instance, in the graph depicted in Figure 4.1, the first two levels from below contain the elements of the information bearing index expressions set.

The last point to explain is how to calculate the probability of a node I given evidence J. For instance, what is $P(\text{cruel} \circ \text{murder} \mid \text{murder } \mathit{by} \text{ Brutus})$? Here, we consider *murder by Brutus* to be evidence, e.g., we know that object d is about *murder by Brutus*, the probability of this index expression is 1. Consequently all the subindex expression of the evidence are true, since, for instance, knowing that an object is about *murder by Brutus* is true, it is natural to assume that the object is about *murder* is true. So, the calculation is then,

$$\begin{aligned} P(\text{Cr} \circ \text{Mu} \mid \text{Mu } \mathit{by} \text{ Br}) &= \gamma_{\text{Cr} \circ \text{Mu}}(\text{Cr} \circ \text{Mu} \mid Cr \wedge Mu) \times \gamma_{\text{Cr}}(Cr) \times \gamma_{\text{Mu}}(Mu) \\ &= \mathbf{P}(\circ) \times \mathit{freq}(Cr) \times 1. \end{aligned}$$

See [23] and [77] for a detailed presentation of the method.

Definition 4.15 Let $\mathcal{D}$ be a set of documents and d be a document with $d \in \mathcal{D}$. Furthermore, let (G, Γ) be a belief network such that the set of probabilistic variables V_G represents a set of index expressions with $q \in G$ and $\chi(d) \subseteq G$, where $\chi(d)$ represents the descriptor set of document d and q a query. The *IEBN aboutness decision* is defined as follows:

$$\models_{\mathrm{IE}_m} d \text{ about } q \text{ if and only if } \exists_{x \in \chi(d)}[P(x \mid q) > P(x)].$$

Now, the intuition behind this definition for information retrieval is as follows. If q increases our belief in descriptors of $\chi(d)$, then it is assumed that d is about q.

Translation

The question is what the index expressions of an IEBN represent. In a typical belief network, nodes are events. We propose that in the IEBN the nodes are situations and the dependencies are aboutness dependencies. The descriptor set of a document $\chi(d)$ which is a set of index-expressions, represents also all those index expressions which are informationally contained in the set. We can say that the document representation is closed under information containment. First we define an *index expression infon language*, which is a subset of $\mathcal{I}_{Idx}(\mathcal{T})$ given in Definition 3.7 at page 42.

Definition 4.16 Let $\mathcal{P}^+(\mathcal{T})$ be a profon language with only positive profons as given in Definition 3.3 and C be a finite set of connectors $\{c_1, \ldots, c_n\}$. The *index expression infon language* $\mathcal{I}_{IE}(\mathcal{T})$ is defined to be the smallest superset of $\mathcal{P}^+(\mathcal{T})$ such that

- if $a_1, a_2 \in \mathcal{P}^+(\mathcal{T})$ and $c \in C$ then $\langle\langle c, a_1, a_2;\ 1\rangle\rangle \in \mathcal{I}_{IE}(\mathcal{T})$.

The infon language $\mathcal{I}_{IE}(\mathcal{T})$ is a sub-language of $\mathcal{I}_{Idx}(\mathcal{T})$. The following infons are excluded: $\{\langle\langle c, a_1, \ldots, a_n;\ \mathrm{i}\rangle\rangle \mid n > 2 \text{ or } i = 0 \text{ or } a_j \notin \mathcal{P}^+(\mathcal{T})\}$. The language $\mathcal{I}_{IE}(\mathcal{T})$ is finite in contrast with the infinite language $\mathcal{I}_{Idx}(\mathcal{T})$.

The language of situations $\mathcal{S}_{IE}$ is the language $\mathcal{S}(\mathcal{I}_{IE}(\mathcal{T}))$. Furthermore, we assumed situations to be closed under information containment, which is defined by:

Definition 4.17 The information containment relation (denoted by $\rightarrow$) is defined by $\langle\langle c, a_1, a_2;\ 1\rangle\rangle \rightarrow a_1$ and $\langle\langle c, a_1, a_2;\ 1\rangle\rangle \rightarrow a_2$.

So, if $\varphi \in S$ and $\varphi \rightarrow \psi$ then $\psi \in S$. The set equivalence is defined analogously to the one of the aboutness proof system SC_{ps}.

Let $\chi(d) = \{I_1, \ldots, I_n\}$ The function $map_1(\chi(d))$ is defined as follows:

$$\begin{aligned} map_1(\chi(d)) &= \{map_2(i) \mid i \in \wp(I)^+ \text{ and } I \in \chi(d)\} \\ map_2(x) &= \langle\langle c, map_2(x_1), map_2(x_2);\ 1\rangle\rangle \text{ if } x \text{ has the form } x_1\ c\ x_2 \\ map_2(x) &= \langle\langle \mathrm{I}, \mathrm{x};\ 1\rangle\rangle \ \text{ if } x \in \mathcal{T}. \end{aligned}$$

For the query we will not introduce a new map function. The function map_1 is defined with as input a set and q is one index expression. In order to avoid type problems we represent the query q, as a singleton set with one index expression. The mapping functions of $\chi(d)$ and q are then the same.

Example 4.1 Examine the following index expression $\chi(d) =$ '(cruel ∘ murder) of Caesar by Brutus', abbreviated to '(Cr ∘ Mu) of Ca by Br'. The translation of the index expression proceeds as follows:

$$\begin{aligned}
map_1(\chi(d)) &= \{map_2(x) \mid x \in \{\mathrm{Cr} \ \circ \ \mathrm{Mu}, \mathrm{Mu\ of\ Ca}, \mathrm{Cr}, \mathrm{Mu}, \mathrm{Ca}\}\} \\
map_2(\mathrm{Cr} \ \circ \ \mathrm{Mu}) &= \langle\langle \circ, map_2(\mathrm{Cr}), map_2(\mathrm{Mu}); 1\rangle\rangle \\
map_2(\mathrm{Cr}) &= \langle\langle \mathrm{I}, \mathrm{Cr}; 1\rangle\rangle \\
&\vdots \\
map_1(\chi(d)) &= \{\langle\langle \circ, \langle\langle \mathrm{I}, \mathrm{Cr}; 1\rangle\rangle, \langle\langle \mathrm{I}, \mathrm{Mu}; 1\rangle\rangle; 1\rangle\rangle, \ldots, \langle\langle \mathrm{I}, \mathrm{Br}; 1\rangle\rangle\}.
\end{aligned}$$

The function map_1 is not injective. As a counterexample, take $\chi(d_1) = \{I_1, I_2\}$ and $\chi(d_2) = \{I_1\}$ with $\wp(I_1)^+ \supseteq \wp(I_2)^+$. Then, $map_1(\chi(d_1)) = map_1(\chi(d_2))$ but $\chi(d_1) \neq \chi(d_2)$. Another counterexample is the following, $\chi(d_1) = \{t_1 \ c \ t_2 \ c \ t_3 \ c \ t_1 \ c \ t_2\}$ and $\chi(d_2) = \{t_1 \ c \ t_2 \ c \ t_3 \ c \ t_1\}$. Here, also $map_1(\chi(d_1)) = map_1(\chi(d_2))$. However, for both cases, in the index expression belief network it also holds that the index expression aboutness decisions of $\chi(d_1)$ are identical to the decisions of $\chi(d_2)$. The function is surjective, due to fact that situations are finite and closed under information containment. Without this additional requirement we could have the situation $\{\langle\langle c, a_1, a_2; 1\rangle\rangle\}$ without a set $A, A \subseteq G$ such that $map_1(A) = \{\langle\langle c, a_1, a_2; 1\rangle\rangle\}$.

Postulates

The aboutness derivation of the IEBN is based on a probabilistic estimation. It is therefore remarkable that it is still possible to extract aboutness derivation steps.

The underlying aboutness proof system of the IEBN retrieval is denoted by IE_{ps} and is defined as follows.

Definition 4.18 (IEBN Situation Aboutness) The aboutness proof system IE_{ps} is defined to be the triple $\langle \mathcal{L}(\mathcal{I}_{IE}(\mathcal{T}))$, {Singleton Reflexivity},{Set Equivalence,Left Monotonic Union,Symmetry,Strict Composition}$\rangle$.

Comparing the aboutness proof system IE_{ps} with the aboutness proof systems C_{ps} and VC_{ps}, we see that they are identical. They only differ in their input language. By definition we can deduce that $\mathcal{I}_{Basic}(\mathcal{T}) \subseteq \mathcal{I}_{IE}(\mathcal{T})$. The consequence of this will be inspected in Chapter 5.

Theorem 4.6 For two index expressions elements I,J of V_G with the number of nodes in V_G larger than 1:

$$P(I \mid J) > P(J) \Leftrightarrow map_1(\{I\}) \cap map_1(\{J\}) \not\equiv \emptyset.$$

Proof Theorem 4.5 expresses that $P(I)$ depends on the probabilities of the information bearing index expression subset only. We have that,

$$\begin{aligned} P(I) &= \prod_{t_i\ c\ t_j \in \wp(I)^+} \gamma_{t_i c t_j}(t_i\ c\ t_j \mid t_i \wedge t_j) \times \prod_{t_i \in \wp(I)^+} \gamma_{t_i}(t_i) \\ &= \prod_{t_i\ c_k\ t_j \in \wp(I)^+} \mathbf{P}(c_k) \times \prod_{t_i \in \wp(I)^+} \mathit{freq}(t_i). \end{aligned}$$

For $P(I \mid J)$, all subindex expressions of J will have the probability assessment 1. Hence, if t $\in \wp(J)$ then t is a subindex expression of J. Consequently $\gamma_{\mathsf{t}}(t) = 1$, which is more than the initial probability assessment of t, namely, $\mathit{freq}(t)$. If $P(I \mid J)$ depends on $\gamma_{\mathsf{t}}(t)$ then $P(I \mid J) > P(I)$. Now, $P(I \mid J)$ depends on $\gamma_{\mathsf{t}}(t)$ if t $\in \wp(I)^+$ as stated by Theorem 4.5. To conclude we have that t $\in \wp(I)^+$ and t $\in \wp(J)^+$. Then by definition of the function map_1 we have that $\langle\langle$I,t; 1$\rangle\rangle \in map_1(\{I\})$ and $\langle\langle$I,t; 1$\rangle\rangle \in map_1(\{J\})$. This proves the theorem.

□

Based on this theorem we can use the soundness and completeness theorem of C$_{ps}$ and conclude that:

Corollary 4.3

(i) The aboutness proof system IE$_{ps}$ is sound. That is, for all subsets A of G and $D \in \mathcal{D}$ such that $\chi(D) = A$ and $B \in G$: if $\vdash_{\mathrm{IE}_{ps}} map_1(A) \Box\!\!\rightsquigarrow map_1(B)$ then $\models_{\mathrm{IE}_m} D$ about B.

(ii) The aboutness proof system IE$_{ps}$ is complete. That is, for all subsets A of G and $D \in \mathcal{D}$ such that $\chi(D) = A$ and $B \in G$: if $\models_{\mathrm{IE}_m} D$ about B then $\vdash_{\mathrm{IE}_{ps}}$ map_1(A) $\Box\!\!\rightsquigarrow map_1$(B).

Reflection

The fact that the index expression aboutness decision is based on simple overlap between the document and query-representation was noticed by IJdens [77]. He remarked that

> *'If a one-term overlap is enough to consider a document to be relevant, the structure of the document captured by the index expressions is not used in the selection of relevant documents.'*

One suggestion made by IJdens, in order to conclude less aboutness decisions, is to demand that document and query should have at least two terms in common. Or in the framework, by replacing the Singleton Reflexivity with the axiom $\{\varphi,\psi\}\,\Box\!\!\rightsquigarrow\{\varphi,\psi\}$, a kind of binary set equivalence. Another, less ad-hoc, suggestion made, was to adopt the axioms $\{\langle\langle c, \langle\langle \mathrm{x}\rangle\rangle, \langle\langle \mathrm{y}\rangle\rangle;\ 1\rangle\rangle\}\,\Box\!\!\rightsquigarrow\{\langle\langle d, \langle\langle \mathrm{x}\rangle\rangle, \langle\langle \mathrm{y}\rangle\rangle;\ 1\rangle\rangle\}$ with c, d elements of the connector set. A weaker condition could be expressed by adopting the axiom in addition with the following axioms $\{\langle\langle c, \langle\langle \mathrm{x}\rangle\rangle, \langle\langle \mathrm{y}\rangle\rangle;\ 1\rangle\rangle\}\,\Box\!\!\rightsquigarrow\{\langle\langle c, \langle\langle \mathrm{y}\rangle\rangle, \langle\langle \mathrm{x}\rangle\rangle;\ 1\rangle\rangle\}$. Taking the last two axioms together, it describes the deduction that in an extended IEBN-model 'information $\circ$ retrieval' and 'retrieval of information' should be considered to be about each other. This criterion is named *Contextual Preselection* [77].

4.5 Boolean retrieval

The model

In boolean retrieval the representation of the documents also consists of a set of terms originating from a descriptor set $\mathcal{T}$. The request is specified as a formula. These formulae are constructed from the descriptor set $\mathcal{T}$ using the logical connectives $\vee$,$\wedge$, and $\neg$. A formula may contain the negation symbol $\neg$, expressing for example that the user wants documents that are not about a certain keyword.

The boolean retrieval inference mechanism is based on the notion of derivation of classical logic to which the Closed World Assumption (CWA) rule, introduced by Reiter [120], is added. We transform Reiter's CWA rule for the purpose of boolean retrieval thus:

Definition 4.19 (Reiter 1978) The closure of a theory D, denoted by $CWA(D)$, is the theory $D \cup \{\neg t : D \nvdash t \text{ and } t \in \mathcal{T}\}$. The set of all theorems derivable from D by CWA is identified with the set of all formulae classically derivable from $CWA(D)$.

Definition 4.20 Let $\mathcal{D}$ be a document-base and $d \in \mathcal{D}$ some document. Furthermore, suppose that $\mathcal{T}$ is some finite set of basic information items (descriptors) such that $\chi(d)$ is a subset of $\mathcal{T}$, where $\chi(d)$ represents the descriptor set of document d. Let the query q be a logical formula constructed from the descriptor set $\mathcal{T}$ using the logical connectives $\vee$,$\wedge$, and $\neg$. We define the *boolean aboutness decision* as follows:

$$\models_{\mathrm{B}_m} d \text{ about } q \text{ if and only if } CWA(\chi(d)) \vdash q.$$

Note that the set $\mathcal{T}$ is used as a set of propositional constants. The truth-value of propositional constant t represents the occurrence of the keyword t in a document. The next step is to define boolean retrieval in terms of our framework. Translating the CWA can proceed in two different ways: (1) one can extend the document representation with the negation of all the terms which are not contained in the representation, or (2) one can add a postulate that expresses the CWA in terms of a rule.

4.5.1 Boolean model I

Let us first consider the option where the document representation is extended with the negation of all terms which are not contained in the representation.

Translation boolean model I

In order to translate boolean retrieval to the framework we define a *boolean infon language* $\mathcal{I}_B(\mathcal{T})$ similar, although not equal, to the boolean infon language $\mathcal{I}_{Bl}(\mathcal{T})$ given in Definition 3.5. Here, we do not introduce the logical connectives $\wedge$ and $\neg$ as elements of the set of relations Rel. We present the connective $\wedge$ using the conjunction operator between situations, and the negation is handled by the polarity of the infons.

Definition 4.21 The *boolean infon language* $\mathcal{I}_B(\mathcal{T})$ is the infon language $\mathcal{I}(\mathcal{P}(\mathcal{T}), \{\vee\}, \emptyset)$.

The language of situations $\mathcal{S}_B$ is the language $\mathcal{S}(\mathcal{I}_B(\mathcal{T}))$. The translation of a document d of a given document-base $\mathcal{D}$ into a situation of $\mathcal{S}_B$ is defined as follows:

$$map_1(\chi(d)) = \{\langle\langle \mathrm{I},t;\ 1\rangle\rangle \mid t \in \chi(d) \text{ and } t \in \mathcal{T}\} \cup \{\langle\langle \mathrm{I},t;\ 0\rangle\rangle \mid t \notin \chi(d) \text{ and } t \in \mathcal{T}\}.$$

Example 4.2 Examine the following two documents: d_1 contains the information that '**C**aesar **l**ikes **B**rutus' and d_2 that '**A**ntonius **h**ates **B**rutus'. The document descriptor sets can be translated as follows:

$$\begin{aligned}
\mathcal{T} &= \{C, L, B, A, H\} \\
\chi(d_1) &= \{C, L, B\} \\
\chi(d_2) &= \{A, H, B\} \\
S_{d_1} &= \{\langle\langle \mathrm{I,C;\ 1}\rangle\rangle, \langle\langle \mathrm{I,L;\ 1}\rangle\rangle, \langle\langle \mathrm{I,B;\ 1}\rangle\rangle, \langle\langle \mathrm{I,A;\ 0}\rangle\rangle, \langle\langle \mathrm{I,H;\ 0}\rangle\rangle\} \\
S_{d_2} &= \{\langle\langle \mathrm{I,C;\ 0}\rangle\rangle, \langle\langle \mathrm{I,L;\ 0}\rangle\rangle, \langle\langle \mathrm{I,B;\ 1}\rangle\rangle, \langle\langle \mathrm{I,A;\ 1}\rangle\rangle, \langle\langle \mathrm{I,H;\ 1}\rangle\rangle\}.
\end{aligned}$$

Without loss of generality we require query formulae to be in *conjunctive normal form*. A formula ϕ is said to be in conjunctive normal form if and only if ϕ is of the form $(\phi_1 \vee \ldots \vee \phi_j) \wedge \ldots \wedge (\phi_k \vee \ldots \vee \phi_m)$ with ϕ_i either a propositional constant t representing a keyword or the negation $\neg t$ of a propositional constant. The negation of a term is modelled as a negative profon ($\langle\langle \mathrm{I,t;0}\rangle\rangle$). The disjunction of two formulae is translated with an infon of the kind $\langle\langle \vee, \phi_1, \ldots, \phi_n;\ 1\rangle\rangle$.

A translation function map_2 from the set of boolean formulae to the set of situations is given as follows:

$$\begin{aligned}
map_2(\phi_1 \wedge \ldots \wedge \phi_n) &= \{map_2(\phi_1)\} \cup \ldots \cup \{map_2(\phi_n)\}\\
map_2(\phi_1 \vee \ldots \vee \phi_n) &= \langle\langle \vee, map_2(\phi_1), \ldots, map_2(\phi_n);\ 1\rangle\rangle\\
map_2(t) &= \langle\langle \mathrm{I}, t;\ 1\rangle\rangle \text{ with } t \in \mathcal{T} \text{ a propositional constant}\\
map_2(\neg t) &= \langle\langle \mathrm{I}, t;\ 0\rangle\rangle \text{ with } t \in \mathcal{T} \text{ a propositional constant.}
\end{aligned}$$

Example 4.3 Consider the boolean query $(C \vee \neg A) \wedge (H \vee B)$, which represents the information-need 'I want all the documents which contain information about "Caesar" or do not contain information about "Antonius", and contain information about "Hate" or "Brutus"'. Using the map_2-function this query is translated as follows:

$$\begin{aligned}
map_2((C \vee \neg A) \wedge (H \vee B)) &= \{map_2(C \vee \neg A)\} \cup \{map_2(H \vee B)\}\\
map_2(C \vee \neg A) &= \langle\langle \vee, map_2(C), map_2(\neg A);\ 1\rangle\rangle\\
map_2(H \vee B) &= \langle\langle \vee, map_2(H), map_2(B);\ 1\rangle\rangle\\
map_2(C) &= \langle\langle \mathrm{I}, \mathrm{C};\ 1\rangle\rangle\\
\vdots &= \vdots\\
map_2((C \vee \neg A) \wedge (H \vee B)) &= \{\langle\langle \vee, \langle\langle \mathrm{I}, \mathrm{C};\ 1\rangle\rangle, \langle\langle \mathrm{I}, \mathrm{A};\ 0\rangle\rangle;\ 1\rangle\rangle,\\
&\quad\ \langle\langle \vee, \langle\langle \mathrm{I}, \mathrm{H};\ 1\rangle\rangle, \langle\langle \mathrm{I}, \mathrm{B};\ 1\rangle\rangle;\ 1\rangle\rangle\}.
\end{aligned}$$

To sum up, we have introduced two translation functions, namely map_1 which translates the keywords of the descriptor set of the document into a set of profons; and map_2 translates boolean formulae into a set of infons. The map_1 function is injective. However, it is not surjective. For instance, the situation S defined as $\{\langle\langle \mathrm{I}, \mathrm{t};\ 1\rangle\rangle, \langle\langle \mathrm{I}, \mathrm{t};\ 0\rangle\rangle\}$ does not have a document description for which $map_1(\chi(d)) = S$. In order to propose a sub-domain in which the function map_1 is surjective, we define a *boolean document situation*.

Definition 4.22 A situation $S \in \mathcal{S}_B$ is called a *boolean document situation* if and only if it satisfies the following conditions:

(a) for all $t \in \mathcal{T}$: $\langle\langle \mathrm{I}, \mathrm{t};\ 0\rangle\rangle \in S$ or $\langle\langle \mathrm{I}, \mathrm{t};\ 1\rangle\rangle \in S$,
(b) if $\langle\langle \mathrm{I}, \mathrm{t};\ i\rangle\rangle \in S$ then $\langle\langle \mathrm{I}, \mathrm{t};\ 1 \Leftrightarrow i\rangle\rangle \notin S$ for $i \in \{0, 1\}$,
(c) if $\phi \in S$ then $\phi \in \mathcal{P}$.

The first condition requires that for all elements of $\mathcal{T}$ there is a positive profon $\langle\langle \mathrm{I}, \mathrm{t};\ 1\rangle\rangle$ or a negative profon $\langle\langle \mathrm{I}, \mathrm{t};\ 0\rangle\rangle$ in the boolean document situation. The second condition requires that if the positive infon of descriptor t is an element of the situation, the negative profon is not, and vice versa. The last condition states that only profons are elements of the boolean document situation.

Claim 4.4 If S and T are boolean document situations and $T \equiv S \cup U$, then $U \subseteq S$.

Proof Reflecting on the conditions of Definition 4.22 we have that: Condition (a) expresses that the number of elements of a boolean document situation is equal with the number of descriptors in $\mathcal{T}$. Consequently, the number of elements of S and T are equal. Since $T \equiv S \cup U$, we have that $U \subseteq S$.

□

The function map_2 is not injective. In case of syntactically different yet logically equivalent formulae the result will be identical situations. Take for example the formulae $t \wedge t$ and t, for which holds that $map_2(t \wedge t) = map_2(t) = \{\langle\langle \mathrm{I},\mathrm{t};\ 1\rangle\rangle\}$. Note that this is not the case for all logical equivalent formulae: for instance, $map_2(t \vee s) \not\equiv map_2(s \vee t)$. The function map_2 is surjective: for every situation there exists a formula.

Postulates

Given the functions map_1 and map_2 one can define an aboutness proof system $\mathrm{B1}_{ps}$ using the output of these two functions.

Definition 4.23 (Boolean Situation Aboutness I) The aboutness proof system $\mathrm{B1}_{ps}$ is defined to be the triple $\langle\mathcal{L}(\mathcal{I}_B(\mathcal{T}))$,{Reflexivity},{Set Equivalence,Cut,Left Monotonic Union,∨-Right Monotonic Composition}⟩.

The rule ∨-Right Monotonic Composition intuitively allows us to deduce that if a situation is about "Caesar" it is also about "Caesar" or "Brutus". Given the conclusion that situation S is about the singleton set $\{\phi\}$, one is able to deduce that S is about the infon $\langle\langle\vee,\phi,\psi;\ 1\rangle\rangle$. One could also derive that S is about $\langle\langle\vee,\phi,\langle\langle\vee,\varphi,\psi;\ 1\rangle\rangle;\ 1\rangle\rangle$. In order to deduce aboutness of infons representing that relation ∨ holds between more than two objects, we consider the infon $\langle\langle\vee,\varphi_1,\ldots,\varphi_k,\psi;\ 1\rangle\rangle$ to be identical to the infon $\langle\langle\vee,\varphi_1,\ldots,\varphi_k,\psi_1,\ldots,\psi_n;\ 1\rangle\rangle$ if $\psi = \langle\langle\vee,\psi_1,\ldots,\psi_n;\ 1\rangle\rangle$.

For instance, $\langle\langle\vee,\phi,\langle\langle\vee,\varphi,\psi;\ 1\rangle\rangle;\ 1\rangle\rangle = \langle\langle\vee,\phi,\varphi,\psi;\ 1\rangle\rangle$. Furthermore we consider that a permutation of the objects in infons of the form $\langle\langle\vee,\varphi_1,\ldots,\varphi_n;\ 1\rangle\rangle$ does not change the information carried by the object. So, for instance $\langle\langle\vee,\phi,\psi;\ 1\rangle\rangle = \langle\langle\vee,\psi,\phi;\ 1\rangle\rangle$. The definition of set equivalence can then be defined as:

$$S \equiv T =^{def} \phi \in S \Leftrightarrow \psi \in T \text{ and } \phi = \psi.$$

Theorem 4.7 The aboutness proof system $\mathrm{B1}_{ps}$ is sound. That is, for all subsets A of $\mathcal{T}$ and $D \in \mathcal{D}$ such that $\chi(D) = A$ and for all logical formulae B in conjunctive normal form constructed from the descriptor set $\mathcal{T}$ using the connectives ∨,∧, and ¬: if $\vdash_{\mathrm{B1}_{ps}} map_1(A) \,\Box\!\!\rightsquigarrow map_2(B)$ then $\models_{\mathrm{B}_m} D$ about B.

Proof First we prove the soundness of the axiom Reflexivity. Secondly we prove the soundness of all the rules of $\mathrm{B1}_{ps}$. This enable us to conclude that $\mathrm{B1}_{ps}$ is sound with respect to B_m.

- The axiom Reflexivity is sound. We have to show that, if $map_1(A) \equiv S$ and $map_2(B) \equiv S$ then $CWA(A) \vdash B$. Due to the fact that $map_1(A) \equiv map_2(B)$ we have no disjunctions in B, as by the definition of the function map_1, infons of the type $\langle\langle \vee, \varphi_1, \ldots, \varphi_n;\ \mathrm{i}\rangle\rangle$ are not in $map_1(A)$, and consequently, not in $map_2(B)$. Let $A = \{a_1, \ldots a_k\}$ with $a_i \in \mathcal{T}$ and $B = (b_1 \wedge \ldots \wedge b_l)$ with b_i a literal. We have that $\langle\langle \mathrm{I}, a_i;\ \mathrm{p}\rangle\rangle \in map_1(A) \Leftrightarrow \langle\langle \mathrm{I}, b_i;\ \mathrm{p}\rangle\rangle \in map_2(B)$ for $1 \leq i \leq \mid map_1(A) \mid$ and $\mathrm{p} = \{0, 1\}$. If $\langle\langle \mathrm{I}, a_i;\ 1\rangle\rangle \in map_1(A)$ then $a_i \in A$, and $a_i \in B$. Furthermore, if $\langle\langle \mathrm{I}, a_i;\ 0\rangle\rangle \in map_1(A)$ then $a_i \notin A$ and $\neg a_i \in B$. By the definition of the function CWA we have that $CWA(A) \vdash B$. This proves the soundness of the axiom.
- The rule Set Equivalence is sound. As with the soundness proofs of SC_{ps} and C_{ps} we restrict ourselves to proving the soundness of the Left Set Equivalence rule. Given that $map_1(A) \equiv map_1(B)$ and $map_1(A) \Box\!\!\leadsto map_2(C)$, which implies that $A \equiv B$ and $CWA(A) \vdash C$. We have to inspect whether the conclusion $CWA(B) \vdash C$ is sound. By the definition of CWA trivially given that $A \equiv B$ if $CWA(A) \vdash C$ then $CWA(B) \vdash C$. This proves the soundness of the Set Equivalence rule.
- The rule Left Monotonic Union is sound. Given that $S \Box\!\!\leadsto T$ is a sound premise, one has to prove that $S \cup U \Box\!\!\leadsto T$ is also sound. Let $S \equiv map_1(A)$, $T \equiv map_2(B)$, and $S \cup U \equiv map_1(C)$. If $S \Box\!\!\leadsto T$ is sound, then $CWA(A) \vdash B$. Now, we have to inspect whether $CWA(C) \vdash B$ under the assumption that $map_1(C) \equiv S \cup U$. Note that due to the soundness of premise $map_1(A)$ is a boolean situation.

  Referring to Claim 4.4, the reader can check easily that $map_1(C)$ is a boolean situation if and only if $S \cup U \equiv S$. Then given that $S \Box\!\!\leadsto T$ is sound, the conclusion $S \cup U \Box\!\!\leadsto T$ is sound, since we proved that the rule Set Equivalence is sound and $S \cup U \equiv S$. This proves the soundness of the rule.
- The proof of the soundness of the Cut rule proceeds as follows. Given that $S \cup T \Box\!\!\leadsto U$ and $S \Box\!\!\leadsto T$ are sound premises we have to prove that $S \Box\!\!\leadsto U$ is a sound conclusion. The argument is similar with the one of the proof of the soundness of Left Monotonic Union. Due to the fact that $S \cup T \Box\!\!\leadsto U$ and $S \Box\!\!\leadsto T$ are sound premises, $S \cup T$ and S are document situations and as shown by the claim $S \equiv S \cup T$, therefore $S \Box\!\!\leadsto U$ is a sound conclusion, which proves the soundness of the Cut rule.
- Finally we have to prove the soundness of the $\vee$-Right Monotonic Composition. If $S \Box\!\!\leadsto \{\varphi\}$ is a sound premise, the conclusion $S \Box\!\!\leadsto \{\langle\langle \vee, \varphi, \psi;\ 1\rangle\rangle\}$ is sound. Let $S \equiv map_1(A)$, $map_2(B) \equiv \{\varphi\}$ and $map_2(C) \equiv \{\langle\langle \vee, \varphi, \psi;\ 1\rangle\rangle\}$. By definition of the function map_2 we have that C is logically equivalent with $B \vee D$, for $map_2(D) = \{\psi\}$. Since $CWA(A) \vdash B$ then $CWA(A) \vdash B \vee D$ is sound, the rule

$\vee$-Right Monotonic Composition is sound.

$\Box$

At first sight the rule Left Monotonic Union seems to be a rule that does not hold in the aboutness proof system that models an IR model with a CWA. The following deduction seems to be sound:

$$\frac{\{\langle\langle \mathrm{I}, t;\ 0\rangle\rangle\} \,\Box\!\!\rightsquigarrow \{\langle\langle \mathrm{I}, t;\ 0\rangle\rangle\}}{\{\langle\langle \mathrm{I}, t;\ 1\rangle\rangle\} \cup \{\langle\langle \mathrm{I}, t;\ 0\rangle\rangle\} \,\Box\!\!\rightsquigarrow \{\langle\langle \mathrm{I}, t;\ 0\rangle\rangle\}}\ \mathsf{LMU}$$

So, assuming the above deduction to be sound, a situation could be about $\{\langle\langle \mathrm{I}, t;\ 1\rangle\rangle\}$ and $\{\langle\langle \mathrm{I}, t;\ 0\rangle\rangle\}$. However, notice that this situation does not result in an document situation. If $\vdash_{\mathrm{B1}_{ps}} map_1(A) \,\Box\!\!\rightsquigarrow map_2(B)$ then $map_1(A)$ is a boolean document situation. Thus, if we have that $S \,\Box\!\!\rightsquigarrow T$, with S is not a boolean document situation due to the fact that some profons (positive or negative) profons are absent, we can add with the rule Left Monotonic Union the absent profons to S in order to achieve a document situation S'.

Theorem 4.8 The aboutness proof system $\mathrm{B1}_{ps}$ is complete. That is, for all subsets A of $\mathcal{T}$ and $D \in \mathcal{D}$ such that $\chi(D) = A$ and for all logical formulae B in conjunctive normal form constructed from the descriptor set $\mathcal{T}$ using the connectives $\vee$,$\wedge$, and $\neg$: if $\models_{\mathrm{B}_m} D$ about B then $\vdash_{\mathrm{B1}_{ps}} map_1(A) \,\Box\!\!\rightsquigarrow map_2(B)$.

Proof We have to show that if $\models_{\mathrm{B}_m} CWA(A)$ about B then $\vdash_{\mathrm{B1}_{ps}} map_1(A) \,\Box\!\!\rightsquigarrow map_2(B)$. Due to the fact that B is in conjunctive normal form, B can be presented as $B = (b_{11} \vee \ldots \vee b_{1s}) \wedge (b_{21} \vee \ldots \vee b_{2t}) \wedge \ldots \wedge (b_{n1} \vee \ldots \vee b_{nz})$ with b_{ij} a literal. If D is about B, then $CWA(A) \vdash B$. Observe that

$$\begin{array}{lcl} CWA(A) \vdash B & \Leftrightarrow & (CWA(A) \vdash b_{11} \text{ or } \ldots \text{ or } CWA(A) \vdash b_{1s}) \\ & & \text{and} \\ & & \vdots \\ & & \text{and} \\ & & (CWA(A) \vdash b_{n1} \text{ or } \ldots \text{ or } CWA(A) \vdash b_{nz}). \end{array}$$

Next, we inspect $CWA(A) \vdash b_{ij}$ with b_{ij} a literal. The reader may verify the following, if b_{ij} is a positive literal:

$$\begin{array}{rcl} CWA(A) \vdash b_{ij} & \Leftrightarrow & b_{ij} \in A, \\ CWA(A) \vdash \neg b_{ij} & \Leftrightarrow & b_{ij} \notin A. \end{array}$$

By Definition 4.19 $CWA(A) = A \cup \{\neg t : A \not\vdash t \text{ and } t \in \mathcal{T}\}$. Due to the fact that A is a set of positive literals, $A \vdash t$ if and only if $t \in A$. Consequently $A \not\vdash t$ if and only if

$t \notin A$. Now given b_{ij} is a positive literal, we have to prove that if $CWA(A) \vdash b_{ij}$ then $\vdash_{\mathrm{B1}_{ps}} map_1(A) \Box\!\!\rightsquigarrow map_2(b_{ij})$ and if $CWA(A) \vdash \neg b_{ij}$ then $\vdash_{\mathrm{B1}_{ps}} map_1(A) \Box\!\!\rightsquigarrow map_2(\neg b_{ij})$. Consider the first case $CWA(A) \vdash b_{ij}$ implies that $b_{ij} \in A$. So, $map_1(A) \supseteq map_2(b_{ij})$. Using Reflexivity and Left Monotonic Union, we can conclude that $\vdash_{\mathrm{B1}_{ps}} map_1(A) \Box\!\!\rightsquigarrow map_2(b_{ij})$. Otherwise, if $CWA(A) \vdash \neg b_{ij}$ then $b_{ij} \notin A$, thus $\langle\langle \mathrm{I}, b_{ij};\ 0\rangle\rangle \in map_1(A)$ and $map_2(\neg b_{ij}) = \{\langle\langle \mathrm{I}, b_{ij};\ 0\rangle\rangle\}$. Here we can also deduce that $map_1(A) \Box\!\!\rightsquigarrow map_2(b_{ij})$ using Reflexivity and Left Monotonic Union. Now, continuing the proof, the deduction 'if $CWA(A) \vdash b_{j1}$ or $\ldots$ or $CWA(A) \vdash b_{js}$ then $CWA(A) \vdash b_{j1} \vee \ldots \vee b_{js}$' has its counterpart in the aboutness proof system, if $map_1(A) \Box\!\!\rightsquigarrow map_2(b_{j1})$ or $\ldots$ or $map_1(A) \Box\!\!\rightsquigarrow map_2(b_{js})$ then $map_1(A) \Box\!\!\rightsquigarrow \{\langle\langle \vee, b_{j1}, \ldots, b_{js};\ 1\rangle\rangle\}$ using the rule $\vee$-Right Monotonic Composition. Finally, the deduction 'if $CWA(A) \vdash b_1$ and $\ldots$ and $CWA(A) \vdash b_n$ then $CWA(A) \vdash b_1 \wedge \ldots \wedge b_n$' is governed by the aboutness proof system $\mathrm{B1}_{ps}$ with the rule Context-Free Union as follows: $map_1(A) \Box\!\!\rightsquigarrow map_2(b_1)$ and $\ldots$ and $map_1(A) \Box\!\!\rightsquigarrow map_2(b_n)$ then $map_1(A) \Box\!\!\rightsquigarrow map_2(b_1 \wedge \ldots \wedge b_n)$, which is identical with $map_1(A) \Box\!\!\rightsquigarrow map_2(b_1) \cup \ldots \cup map_2(b_n)$.

□

Reflection

As one may have noticed, $\mathrm{B1}_{ps}$ contains the postulates of the system SC_{ps} in addition with rules for the $\vee$-operator. Next, we inspect the top and bottom elements as defined on page 64.

Proposition 4.5 In the aboutness proof system $\mathrm{B1}_{ps}$ we have that:

(i) The top query of $\mathrm{B1}_{ps}$ is the set $\{map_2(\varphi) \mid \varphi \text{ is a tautology}\}$.

(ii) The top document of $\mathrm{B1}_{ps}$ is the set $\{\langle\langle \mathrm{I}, \mathrm{t};\ \mathrm{p}\rangle\rangle \mid \mathrm{p} \in \{0, 1\} \text{ and } \mathrm{t} \in \mathcal{T}\}$.

(iii) The bottom query of $\mathrm{B1}_{ps}$ is the set $\emptyset$.

(iv) The bottom document of $\mathrm{B1}_{ps}$ is the set $\emptyset$.

Tautology is defined as in classical logic.

Proof

(i) We have to show that, for arbitrary situations $S \in \mathcal{S}_B$, the aboutness formula $S \Box\!\!\rightsquigarrow T$ is an aboutness theorem for all $T \in \mathbf{1}^q_{\mathrm{B1}_{ps}}$. A tautology B in classical logic has the property that for all $A : A \vdash B$. So, assume B is a tautology which is in conjunctive normal form $B = (b_{11} \vee \ldots \vee b_{1s}) \wedge \ldots \wedge (b_{n1} \vee \ldots \vee b_{nz})$ with b_{ij} a literal. This implies that for all A: $A \vdash (b_{11} \vee \ldots \vee b_{1s})$ and $\ldots$ and $A \vdash (b_{n1} \vee \ldots b_{nz})$. Let us inspect a tautology $(b_{11} \vee \ldots \vee b_{1s})$ in isolation. This implies that there is always one or more literals b_{1j} with $1 \leq j \leq s$ for which $A \vdash b_{1j}$. Due to the completeness of the aboutness proof system, we can conclude that there is always a b_{1j} for which $map_1(A) \Box\!\!\rightsquigarrow map_2(b_{1j})$, and consequently $map_1(A) \Box\!\!\rightsquigarrow map_2(\langle\langle \vee, b_{11}, \ldots, b_{1s};\ 1 \rangle\rangle)$. So, one may conclude that $map_1(A) \Box\!\!\rightsquigarrow map_2(B)$.

(ii) We have to show that, for arbitrary situation $T \in \mathcal{S}_B$, the aboutness formula $S \Box\!\!\rightsquigarrow T$ is an aboutness theorem for all $S \in \mathbf{1}^d_{\mathrm{B1}_{ps}}$. First we note that the element of the top document of $\mathrm{B1}_{ps}$ is a situation S that contains all profons of the the language $\mathcal{I}_B(\mathcal{T})$. Assume T is a situation such that $U \cup V \equiv T$, with U the profons of T and V is a situation with infons of the form $\langle\langle \vee, \varphi_1, \ldots, \varphi_n;\ 1 \rangle\rangle$. Furthermore assume $W \equiv S \backslash U$ then, it is provable from $U \Box\!\!\rightsquigarrow U$, that $U \cup W \Box\!\!\rightsquigarrow U$. Using Set Equivalence one determines that $S \Box\!\!\rightsquigarrow U$. Now, we have to prove that $S \Box\!\!\rightsquigarrow V$, which allows us to conclude that $S \Box\!\!\rightsquigarrow U \cup V$, and using Set Equivalence one determines that $S \Box\!\!\rightsquigarrow T$. The rule $\vee$-Right Monotonic Composition states that given $S \Box\!\!\rightsquigarrow \{\varphi\}$ one may conclude that $S \Box\!\!\rightsquigarrow \{\langle\langle \vee, \varphi, \psi;\ 1 \rangle\rangle\}$. So, for each element of V we have to inspect if one object a_i of the infon $\langle\langle \vee, a_1, \ldots, a_n;\ 1 \rangle\rangle$ is about S. Recursively, we could proceeds as follows: $S \Box\!\!\rightsquigarrow \{\langle\langle \vee, a_1, \ldots, a_n;\ 1 \rangle\rangle\}$ if and only if there is a a_i with $1 \leq i \leq n$ such that $S \Box\!\!\rightsquigarrow \{a_i\}$. If a_i is a profon, then we could prove that $S \Box\!\!\rightsquigarrow \{a_i\}$, using Reflexivity and Set Equivalence analogously as we did for the situation V. Otherwise, we continue recursively on the structure of a_i. This recursion on the structure will determine on a specific profon φ. Here, one can analogously prove that $S \Box\!\!\rightsquigarrow \{\varphi\}$.

(iii) Since $\mathbf{1}^d_{\mathrm{B1}_{ps}}$ is not empty, Proposition 4.2 suffices to conclude that $\mathbf{0}^q_{\mathrm{B1}_{ps}}$ is empty.

(iv) Since $\mathbf{1}^q_{\mathrm{B1}_{ps}}$ is not empty, Proposition 4.2 suffices to conclude that $\mathbf{0}^d_{\mathrm{B1}_{ps}}$ is empty. □

Note that since we do not have a surjective map function, we can not use Proposition 4.1 to derive the top and bottom elements of the model B_m.

Proposition 4.6 In the IR model B_m we have that:

(i) The top query of B_m is the set $\{\varphi \mid \varphi \text{ is a tautology}\}$.
(ii) The bottom query of B_m is the set $\{\varphi \mid \varphi \text{ is a contradiction }\}$.
(iii) The top document of B_m is the set $\emptyset$.
(iv) The bottom document of B_m is the set $\emptyset$.

Tautology and contradiction are defined as in classical logic.

Proof

(i) The proof of item (i) is completely analogous to the proof of item (i) of Proposition 4.5.

(ii) There is no document that is about a contradiction. A contradiction B in classical logic has the property that for all $A : A \not\vdash B$. So, assume B is a contradiction then for all $\chi(d)$, $CWA(\chi(d)) \not\vdash B$. Consequently, there is no document retrieved with respect to query B which proves the proposition.

(iii) Since $\mathbf{0}^q_{\mathrm{B}_m}$ is not empty, Proposition 4.2 suffices to conclude that $\mathbf{1}^d_{\mathrm{B}_m}$ is empty.

(iv) Since $\mathbf{1}^q_{\mathrm{B}_m}$ is not empty, Proposition 4.2 suffices to conclude that $\mathbf{0}^d_{\mathrm{B}_m}$ is empty.

□

One may wonder, whether, given φ is a contradiction, $map_2(\varphi)$ is not an element of the bottom query of $\mathrm{B1}_{ps}$. Since Reflexivity is an axiom of $\mathrm{B1}_{ps}$, we have that $map_2(\varphi) \,\Box\!\!\rightsquigarrow map_2(\varphi)$, which contradicts the assumption.

4.5.2 Boolean model II

The presented approach is based on a 'Closed World' indexing: if a document is not indexed with a descriptor, it is assumed to be indexed with the negation of that particular descriptor. This approach includes the following aspects. First, adding new documents with descriptors not yet in the set $\mathcal{T}$ to the document-base requires an update of all the representations of the documents with the negated form of the new descriptors. Furthermore, the fact that boolean retrieval operates under the CWA is, according to Bruza, 'one of the principle clarifications why these models offer ineffective disclosure' [23]. For instance a document indexed with the keywords "killed", "Brutus", and "Caesar", is about the query **not** "murder".

In order to highlight the 'rule' that is governed by the CWA we present an aboutness proof system in which the CWA is incorporated as a rule rather than through the mapping function.

Translation boolean model II

Another approach consists of adopting the CWA in terms of a rule. The boolean infon language $\mathcal{I}_B(\mathcal{T})$ is defined as in the previous boolean aboutness proof system. The translation function of a document representation into a situation is given as follows:

$$map_1(\chi(d)) = \{\langle\langle \mathrm{I}, \mathrm{t};\ 1\rangle\rangle \mid \mathrm{t} \in \chi(d)\}.$$

Here, the definition of a *boolean document situation* is different than the one given in the previous approach.

Definition 4.24 A situation $S \in \mathcal{S}_B$ is called a *boolean document situation* if and only if $S \subseteq \mathcal{P}^+(\mathcal{T})$.

The query transformation is identical to the translation of the first approach.

Postulates

In this case we do not add negative infons to the representation, but instead we add new rules to obtain the same effect. The document boolean situation is a set of positive profons the query boolean situation is an element of $\mathcal{S}_B$ as defined before.

Definition 4.25 (Boolean Situation Aboutness II) The aboutness proof system $\mathrm{B2}_{ps}$ is defined to be the triple $\langle \mathcal{L}_2^{Ext}(\mathcal{I}_B(\mathcal{T})), \{\text{Reflexivity}_1\}, \{\text{Set Equivalence}_1, \text{Left Monotonic Union}_1, \text{Cut}_1, \text{Context-Free Union}_2, \vee\text{-Right Monotonic Composition}_2, \text{Aboutness Inheritance}, \text{Simple Anti-Aboutness}, \text{Closed World Assumption}\}\rangle$.

Here we are using the extended combined aboutness language $\mathcal{L}_2^{Ext}(\mathcal{I}_B(\mathcal{T}))$ as given on page 56. Therefore some of the axioms and rules have to be indexed. For instance the Cut$_1$ appears as follows:

$$\frac{S \,\Box\!\!\rightsquigarrow_1 T \quad S \cup T \,\Box\!\!\rightsquigarrow_1 U}{S \,\Box\!\!\rightsquigarrow_1 U}$$

Let us explain the aboutness proof system to some extent. The Aboutness Inheritance expresses that all aboutness theorems of 1 are aboutness theorems of 2. Note that $\Box\!\!\rightsquigarrow_1$ is defined with the postulates of the aboutness proof system SC_{ps}. This implies that if $S \,\Box\!\!\rightsquigarrow_1 T$, then $S \supseteq T$, which has been proved in Section 4.1.

Furthermore the rule Simple Anti-Aboutness implies that if one is not able to prove $S \,\Box\!\!\rightsquigarrow_1 T$ then one concludes $S \boxtimes\!\!\rightsquigarrow_1 T$. As mentioned in Chapter 3, we believe that this is not a good definition of an anti-aboutness relation. The rule Context-Free Union$_2$ expresses that given the premises $S \,\Box\!\!\rightsquigarrow_2 T$ and $S \,\Box\!\!\rightsquigarrow_2 U$, then one may conclude that $S \,\Box\!\!\rightsquigarrow_2 T \cup U$.

Finally the postulate Closed World Assumption is a rule of the aboutness proof system. The Closed World Assumption rule deserves more attention. We are using an anti-aboutness decision $S \boxtimes\!\!\rightsquigarrow_1 T$ in combination with a preclusion decision to derive an aboutness decision of $S \,\Box\!\!\rightsquigarrow_2 U$. Preclusion is defined here as $\langle\langle \mathrm{I},\mathrm{t};\ 1\rangle\rangle \perp_2 \langle\langle \mathrm{I},\mathrm{t};\ 0\rangle\rangle$ for all $t \in \mathcal{T}$. Note, that this implies that the preclusion relation is not symmetric, i.e., $\langle\langle \mathrm{I},\mathrm{t};\ 0\rangle\rangle \perp_2 \langle\langle \mathrm{I},\mathrm{t};\ 1\rangle\rangle$ is not an axiom. In order to explain the Closed World Assumption rule (given that $S \boxtimes\!\!\rightsquigarrow_1 \{\phi\}$, and that $\phi \perp_2 \psi$, we can deduce that $S \,\Box\!\!\rightsquigarrow_2 \{\psi\}$) let us consider the following example: we can prove that $\{\langle\langle \mathrm{I},\mathrm{Caesar};\ 1\rangle\rangle\}$ is about $\{\langle\langle \mathrm{I},\mathrm{Brutus};\ 0\rangle\rangle, \langle\langle \mathrm{I},\mathrm{Ceasar};\ 1\rangle\rangle\}$ as follows (in abbreviated form):

$$\mathbf{I}\ \frac{\{\langle\langle \mathrm{I},\mathrm{C};\ 1\rangle\rangle\} \boxtimes\!\!\rightsquigarrow_1 \{\langle\langle \mathrm{I},\mathrm{B};\ 1\rangle\rangle\} \quad \langle\langle \mathrm{I},\mathrm{B};\ 1\rangle\rangle \perp_2 \langle\langle \mathrm{I},\mathrm{B};\ 0\rangle\rangle}{\{\langle\langle \mathrm{I},\mathrm{C};\ 1\rangle\rangle\} \,\Box\!\!\rightsquigarrow_2 \{\langle\langle \mathrm{I},\mathrm{B};\ 0\rangle\rangle\}}\ \mathsf{CWA}$$

$$\mathbf{II}\ \frac{\{\langle\langle \mathrm{I},\mathrm{C};\ 1\rangle\rangle\} \,\Box\!\!\rightsquigarrow_1 \{\langle\langle \mathrm{I},\mathrm{C};\ 1\rangle\rangle\}}{\{\langle\langle \mathrm{I},\mathrm{C};\ 1\rangle\rangle\} \,\Box\!\!\rightsquigarrow_2 \{\langle\langle \mathrm{I},\mathrm{C};\ 1\rangle\rangle\}}\ \mathsf{AI}$$

$$\dfrac{\dfrac{\mathbf{I} \qquad \mathbf{II}}{\{\langle\langle \mathrm{I},\mathrm{C};\ 1\rangle\rangle\}\,\Box\!\!\leadsto_2\{\langle\langle \mathrm{I},\mathrm{B};\ 0\rangle\rangle\} \cup \{\langle\langle \mathrm{I},\mathrm{C};\ 1\rangle\rangle\}}\;\mathsf{CFU}}{\{\langle\langle \mathrm{I},\mathrm{C};\ 1\rangle\rangle\}\,\Box\!\!\leadsto_2\{\langle\langle \mathrm{I},\mathrm{B};\ 0\rangle\rangle, \langle\langle \mathrm{I},\mathrm{C};\ 1\rangle\rangle\}}\;\mathsf{SE}$$

Theorem 4.9 The aboutness proof system $\mathrm{B}2_{ps}$ is sound. That is, for all subsets A of $\mathcal{T}$ and $D \in \mathcal{D}$ such that $\chi(D) = A$ and for all logical formulae B in conjunctive normal form constructed from the descriptor set $\mathcal{T}$ using the connectives $\vee$,$\wedge$, and $\neg$ and for $i \in \{1, 2\}$: if $\vdash_{\mathrm{B}2_{ps}} map_1(A)\,\Box\!\!\leadsto_i map_2(B)$ then $\models_{\mathrm{B}_m} D$ about B.

Proof First we prove the soundness of the aboutness theorems of 1. Secondly we prove the soundness of all the other rules of $\mathrm{B}2_{ps}$. This enable us to conclude that $\mathrm{B}2_{ps}$ is sound with respect to B_m.

- In order to prove that $S\,\Box\!\!\leadsto_1 T$ is a sound theorem we can reflect to the soundness proof of the aboutness proof system SC_{ps}. There we proved that if $S\,\Box\!\!\leadsto T$, then $S \supseteq T$. So, if $map_1(A)\,\Box\!\!\leadsto_1 map_2(B)$ then $map_1(A) \supseteq map_2(B)$. Assume that $S \equiv map_1(A)$ and $T \equiv map_2(B)$. By the definition of function map_1 we have that $map_1(A)$ is a boolean situation, thus $S \subseteq \mathcal{P}^+(\mathcal{T})$, and consequently $T \subseteq \mathcal{P}^+(\mathcal{T})$. By the definition of the function map_2 it follows that $B = t_1 \wedge \ldots \wedge t_n$ with t_i a positive literal and $t_i \in A$ for $1 \leq i \leq n$. Obviously, $CWA(A) \vdash B$, which proves the soundness of the aboutness theorems of 1.
- By the preceding considerations the proof of the soundness of the Aboutness Inheritance rule is finished as well. Since, given the sound premise $S\,\Box\!\!\leadsto_1 T$, then trivially $S\,\Box\!\!\leadsto_2 T$ is sound.
- The soundness of the Closed World Assumption can be proved as follows. Given that $S \boxtimes\!\!\leadsto \{\phi\}$ is a sound premise and that $\phi \bot_2 \psi$. Assume that $S \equiv map_1(A)$ and $\phi = \langle\langle \mathrm{I},\mathrm{t};\ 1\rangle\rangle$. The sound premise implies that $\langle\langle \mathrm{I},\mathrm{t};\ 1\rangle\rangle \notin S$. So, now we have to prove that $S\,\Box\!\!\leadsto \{\langle\langle \mathrm{I},\mathrm{t};\ 0\rangle\rangle\}$ is a sound conclusion. Thus, that $CWA(A) \vdash \neg t$. By definition of the function map_1 we have that if $\langle\langle \mathrm{I},\mathrm{t};\ 1\rangle\rangle \notin S$ then $t \notin A$, which allows us to conclude that $CWA(A) \vdash \neg t$.
- In order to prove the soundness of the rule Context-Free Union$_2$, one has to prove that given $S\,\Box\!\!\leadsto_2 T$ and $S\,\Box\!\!\leadsto_2 U$ are sound, $S\,\Box\!\!\leadsto_2 T \cup U$ is sound. Assume that $S \equiv map_1(A)$, $T \equiv map_2(B)$, and $U \equiv map_2(C)$. By the definition of the function map_2 we have that $map_2(B \wedge C) \equiv map_2(B) \cup map_2(C)$. The sound premise implies that $CWA(A) \vdash B$ and $CWA(A) \vdash C$, which allows us to conclude that $CWA(A) \vdash B \wedge C$. This suffices to conclude that the rule Context-Free Union$_2$ is sound.
- The proof of the soundness of the rule $\vee$-Right Monotonic Composition$_2$ proceeds analogously to the one of aboutness proof system $\mathrm{B}1_{ps}$.

□

Theorem 4.10 The aboutness proof system $\mathrm{B2}_{ps}$ is complete. That is, for all subsets A of $\mathcal{T}$ and $D \in \mathcal{D}$ such that $\chi(D) = A$ and for all logical formulae B in conjunctive normal form constructed from the descriptor set $\mathcal{T}$ using the connectives $\vee$,$\wedge$, and $\neg$ and for $i \in \{1, 2\}$:: if $\models_{\mathrm{B}_m}$ D about B then $\vdash_{\mathrm{B2}_{ps}} map_1(A) \Box\!\!\leadsto_i map_2(B)$.

Proof For the first part of the completeness proof proceeds analogously to the one of the aboutness proof system $\mathrm{B1}_{ps}$. We continue the proof after the conclusion that, if b_{ij} is a positive literal:

$$\begin{aligned} CWA(A) \vdash b_{ij} \quad &\Leftrightarrow \quad b_{ij} \in A, \\ CWA(A) \vdash \neg b_{ij} \quad &\Leftrightarrow \quad b_{ij} \notin A. \end{aligned}$$

Then, we have to prove that $map_1(A) \Box\!\!\leadsto_2 map_2(b_{ij})$. Assume that $S \equiv map_1(A)$ and $map_2(b_{ij}) \equiv \{\langle\langle \mathrm{I}, b_{ij};\ 1\rangle\rangle\}$. If $b_{ij} \notin A$, then $\langle\langle \mathrm{I}, b_{ij};\ 1\rangle\rangle \notin S$. Applying the rule Simple Anti-Aboutness we have that $S \boxtimes\!\!\leadsto_1 \{\langle\langle \mathrm{I}, b_{ij};\ 1\rangle\rangle\}$. By the definition of the preclusion-relation we have that $\langle\langle \mathrm{I}, b_{ij};\ 1\rangle\rangle \perp_2 \langle\langle \mathrm{I}, b_{ij};\ 0\rangle\rangle$. Using the rule Closed World Assumption we can conclude that $S \Box\!\!\leadsto \{\langle\langle \mathrm{I}, b_{ij};\ 0\rangle\rangle\}$. In case $b_{ij} \in A$, then $\langle\langle \mathrm{I}, b_{ij};\ 1\rangle\rangle \in S$. Consequently $S \Box\!\!\leadsto_1 \{\langle\langle \mathrm{I}, b_{ij};\ 1\rangle\rangle\}$ and the rule Aboutness Inheritance allows us to conclude that $S \Box\!\!\leadsto_2$ $\{\langle\langle \mathrm{I}, b_{ij};\ 1\rangle\rangle\}$. The proof proceeds analogously to the completeness proof of $\mathrm{B1}_{ps}$.

□

Reflection

The top query set and the bottom query set are identical to the ones of the aboutness proof system $\mathrm{B1}_{ps}$. As mentioned in Chapter 3, the definition of anti-aboutness in terms of *not about* leads to undesirable properties. The fact that a document titled 'Brutus killed Caesar' is not indexed with the keyword "murder" should not imply that this document is anti-about "murder". In case we want to improve the aboutness proof system $\mathrm{B2}_{ps}$ one could suggest to improve the anti-aboutness definition. This can be done by improving the rule Simple Anti-Aboutness. If one is able to define anti-aboutness more precisely, the retrieval results of $\mathrm{B2}_{ps}$ could be more precise. Based on the same intuition one could suggest an extended definition of the preclusion relation.

4.6 Conceptual Graph retrieval

The model

In this section we introduce the logical IR model Elen as presented in the thesis of Chevallet [34] (see also [73, 74, 114]). Chevallet proposes the use of the conceptual graphs formalism to build an operational version of the logical model suggested by van

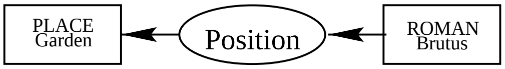
Rijsbergen [124]. The logical model is only a formal framework for designing information retrieval systems involving knowledge and deduction mechanisms. Several logical IR models are already designed to deal with complex information and deductions through an appropriate knowledge representation formalism. For example, within the RIME [18] and MIRTL [101] projects formalisms are used that are based on the notion of Conceptual Dependency and on Terminological Logic, respectively. These projects aim at building operational logical models for IR.

It is in the same direction that Chevallet proposes to use the conceptual graphs formalism to instantiate the logical model. His idea led to the system Elen (géniE logicieL & recherchE d'informatioNs), which is based on an indexing language that uses conceptual graphs. The conceptual graph approach is based on the basic definitions and properties of conceptual graphs as developed by Sowa [137].

Next, we introduce the conceptual graphs in the way in which they are used in the conceptual graph model, which we refer to as Elen. A graph is a representation of information and consists of the following three basic elements:

- ○ concept nodes,
- ○ relation nodes,
- ○ edges between concept and relation nodes.

A concept node, represented graphically by a box, has a *concept type*. This concept type corresponds to a semantic class, e.g., Person,Place, and so on. This concept type has possibly a referent, which corresponds to an instantiation of the class of the concept type. For instance, Brutus, Garden could be referents of the concept types Person and Place respectively. A *relation node*, represented graphically by an oval, has a *relation type* only, which corresponds to a semantic class of relations such as, *ActsOn*, *Position*, and so on. Two concept nodes can be related to each other using a relation node and edges. The edges express in which way the concept nodes are related. The following graph is constructed out of the concepts Place and Roman, the referents Brutus and Garden, and the relation *Position*.

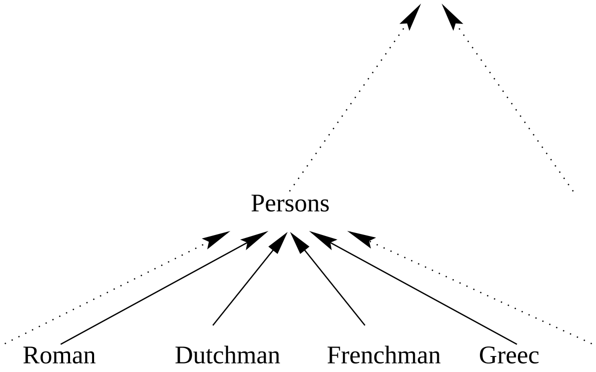


The Roman "Brutus" is in a Place "Garden".

In this graph, the arrows express that the "Garden" is the position of "Brutus" and not "Brutus" is the position of the "Garden". Now we can define the conceptual graph as follows [137]:

Definition 4.26 (Conceptual Graph [Sowa '84]) A conceptual graph is a finite and oriented, bipartite, connected graph of *concepts* and *relations* nodes. In a conceptual graph, concept nodes represent entities, attributes, states and events, and relation nodes represents relations between the concept nodes. The edges show how concept nodes are interconnected by relation nodes.

Besides a graphical representation of information, Sowa [137] introduces a knowledge base. This knowledge base contains a concept and relation type taxonomy. Such a type taxonomy is a lattice structure of types. In this lattice the information is represented that a certain type is semantically included in another type. For instance, it could be the case that ROMAN is semantically included in PERSON. The figure below is an example of a fragment of a concept type taxonomy:



One can use the lattice to induce a partial ordering relation $\leq$. This relation is defined as follows: if X is semantically included in Y according to the lattice then $X \leq Y$. Furthermore, $\leq$ is assumed to be reflexive and transitive. For instance, in the graph depicted above, ROMAN $\leq$ PERSON. We say that type ROMAN is a *restriction* or a *subtype* of PERSON, and that PERSON is a *generalisation* of ROMAN. The knowledge base contains two such lattices, one for the concept types and one for the relation types.

In case of the concept types, the $\leq$ relation can be extended to concept nodes having referents. For example, $\boxed{\text{ROMAN:Brutus}} \leq \boxed{\text{ROMAN}}$, where $\boxed{\text{ROMAN}}$ represents the concept of all Romans, and $\boxed{\text{ROMAN:Brutus}}$ represents the concept of a Roman named "Brutus".

In ELEN documents are represented by graphs. According to Peirce [116] and Sowa [138] it can be argued that it is easier for the user to express her information (sentences) by means of a graph than by using formulae from first order logic. Adopting this point of view, a graph representation is a suitable option to formulate a query.

Furthermore, Chevallet motivated his choice for ELEN by the fact that the conceptual graph formalism can represent all components of an IR system: documents and queries, as well as the general domain knowledge of the document-base.

Next we present how one can index documents by conceptual graphs. In the conceptual graph approach the indexer manually creates a set of conceptual graphs, called the *minimal canonical* graphs. These graphs represent information descriptors, for example, that the concept type PLACE is related by relation *Position* to a concept type PERSON. The reason that they are called canonical is that these graphs are assumed to have a correct informational meaning (from the position of the indexer). A set of minimal canonical graphs from the *minimal canonical base*. This base contains all sufficient and necessary information descriptors, which are needed to represent the document-base.

Furthermore, for each concept type there exists a conformity relation indicating that a referent is a correct instantiation of a concept type. With this relation we can inspect whether "Brutus" is conform to the concept type ROMAN or not. If a concept type X is a generalisation of a concept type Y, and "t" is conform to Y then one may also conclude that "t" is conform to X.

The conformity relation and set of *minimal canonical* graphs are fixed and created manually by a human indexer. Now, new canonical graphs may be generated from existing ones using the following four elementary operators:

(1) Copy: if w is a conceptual graph then a copy u of w is also a conceptual graph.

Figure 4.2: The copy of a graph.

(2) Restriction: a graph is restricted when a concept type or a relation type is replaced by a subtype, or when a referent is replaced by an included set. In Figure 4.3, given the fact that C is a subtype of the concept type A, the left graph can be restricted to the right graph.

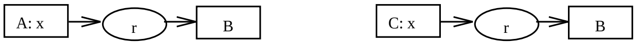


Figure 4.3: The restriction of a graph.

(3) Simplification: when two concepts are linked by two identical relations, then one may be deleted. For instance, in Figure 4.4, the left graph can be simplified to the right graph.

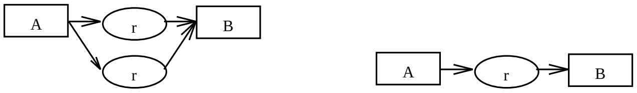


Figure 4.4: The simplification of a graph.

(4) Join: two graphs that have one concept in common, can be joined to form one graph by sharing this common concept. In Figure 4.5 two graphs are joined on their common concept B.

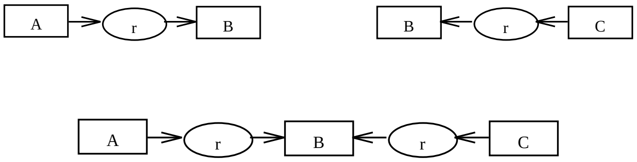


Figure 4.5: The join of two graphs.

These operators can be used to construct new graphs. The result will be a conceptual graph that represents the information of a document or a query q. For the sake of simplicity ELEN adopts the constraint that every document and query are represented by one single graph [137]. Furthermore, ELEN adopts the additional constraint that only dyadic relations are used in the index.

If it is possible to build a graph B starting from the graph A using the four operators, then we can view this graph B as a restriction of graph A. Or, stated differently, that $B \leq A$. This implies that the definition of the $\leq$-relation is extended to graphs. A node with a concept type with or without a referent could be seen as a graph, therefore we could say that $\leq$ is a relation on conceptual graphs. In this perspective Sowa defined $\leq$ as follows: if the graph B is the result of using the four operators starting from graph A then $B \leq A$. Note that in this definition, we still have that a concept type with a restriction is a subtype of the same concept type without a referent, due to the restriction operator on graphs.

The $\leq$ relation defined on conceptual graphs is of prime importance in ELEN. A document d indexed by a conceptual graph $\chi(d)$ is about a query represented by a conceptual graph q, if and only if $\chi(d) \leq q$, i.e., the information contained in graph q is also contained in graph $\chi(d)$. One can say that the relation $\leq$ plays the role of the deduction connective in the logical model.

Sowa [137] introduced a projection operator that makes it clear whether a graph is a specialisation of another graph or not.

Definition 4.27 A conceptual graph H *is projected on* a graph G if and only if there consists a subgraph G' of G that satisfies the following conditions:

(i) The conceptual relations in G' and H are identical.

(ii) The concepts $C_1, \ldots, C_n$ of G' are specialisations of the corresponding concepts $D_1, \ldots, D_n$ of H.

(iii) If a relation R links two concepts D_i and D_j in H, then it also links the concepts C_i and C_j in G'.

Sowa proves that if a conceptual graph G is a specialisation of H, there must exist a projection of H on G. Mugnier [102] shows the converse, e.g., that if there is a projection of H on G, then $G \leq H$. This shows that the projection operator may be viewed as the basic retrieval operator: retrieving documents that imply query q is equivalent to retrieving documents that contain a projection of q.

Definition 4.28 Let $\mathcal{D}$ be a document-base and d a document with $d \in \mathcal{D}$. Furthermore, let $\mathcal{G}$ be a set of conceptual graphs, with q and $\chi(d) \in \mathcal{G}$, where $\chi(d)$ is the representation of document d and q is a query. The *conceptual graph aboutness decision* is defined as follows:

$$\models_{\mathrm{CG}_m} d \text{ about } q \text{ if and only if } \chi(d) \leq q.$$

Let us consider an example of a conceptual graph aboutness decision as given in [113]. A user wants to retrieve all documents dealing with ‘a UNIX command that searches for an object in a structure’. In Figure 4.6 two conceptual graphs are depicted. The query is formulated as a conceptual graph q. The document is represented as a conceptual graph d. This document is a manual of the UNIX command ‘grep’. In this figure, the subgraph of d, which contains darkened nodes corresponds to the projection of q. Note that in this projection, the concepts File and Expression of $\chi(d)$ are restrictions of the concepts Structure and Object of q, respectively. One could verify that q is indeed projected on graph d (or alternatively, that d is a specialisation of q) and therefore retrieved.

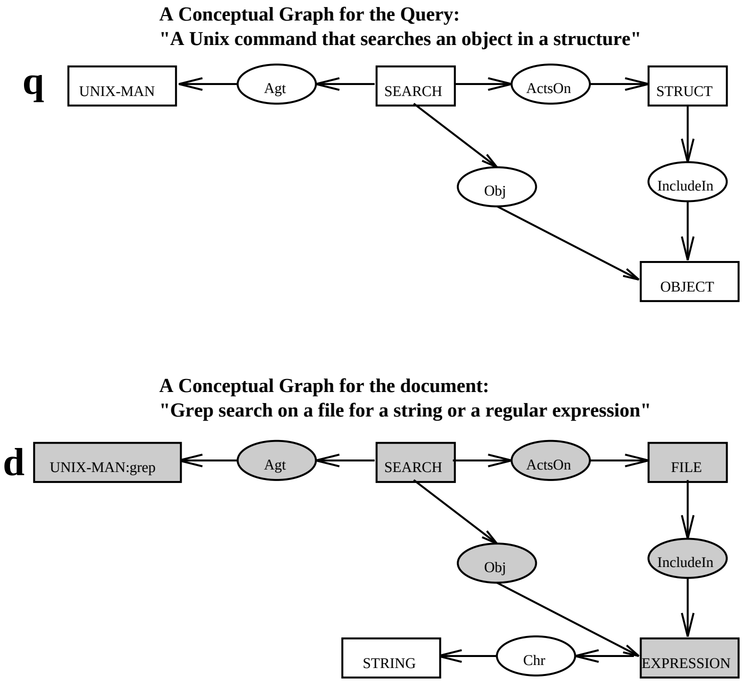

A Conceptual Graph for the Query:
"A Unix command that searches an object in a structure"
q
UNIX-MAN
Agt
SEARCH
ActsOn
STRUCT
Obj
IncludeIn
OBJECT
A Conceptual Graph for the document:
"Grep search on a file for a string or a regular expression"
d
UNIX-MAN:grep
Agt
SEARCH
ActsOn
FILE
Obj
IncludeIn
STRING
Chr
EXPRESSION

Figure 4.6: Document d is *conceptual graph about* query q.

Chevallet [34] noticed that the join operator can be used in three different ways: (i) join two common concept nodes belonging to the same graph, termed *internal join.* (ii) join two common concept nodes belonging to two distinct graphs, termed *external join.* (iii) join all common concept nodes and simplify if possible afterwards, termed *maximal join.*

Consider the following sequence of operators on a graph $G1$:

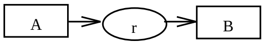


Figure 4.7: Graph G1.

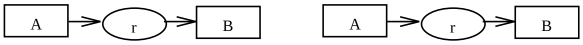


Figure 4.8: After the copy operator.

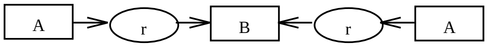


Figure 4.9: External join on common concept node B (Graph G2).

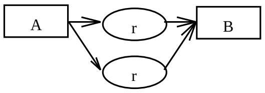


Figure 4.10: Internal join on common concept node A (Graph G3).

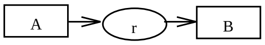


Figure 4.11: Simplification (Graph G1).

Here, we have that $G1 \leq G3$ and $G3 \leq G1$. Hence, $G1$ is a specialisation of $G3$, and $G3$ is a specialisation of $G1$ which implies that the information of $G1$ is identical to the information of $G3$. In case we want to represent the information of a document as a graph precisely, it is important to be aware that a document indexed with graph $G1$ and a document indexed by a graph $G3$ is threated identically by the system. Therefore Chevallet introduced the notion of *normalised graphs.* A graph is *normalised* if no non-empty sequence of simplifications, internal joins or specialisations can be applied to this graph to yield an equivalent graph. In ELEN only normalised graphs are used as representation of the documents. Furthermore, ELEN uses only the maximal join as the join operator, which directly leads to normalised graphs.

Translation

The next step is the translation of conceptual graphs to situations. One of Sowa's important statements about conceptual graphs is that they can be associated to first-order logical formulae through a transformation function (see beside [137], [12] for an in-depth study of the relation between conceptual graphs and logic). We believe, however, that for using conceptual graphs in information retrieval, we need a transformation to an information theory instead of to a truth theory. This belief is explained in Chapter 2.

A conceptual graph carries information, as such it can be seen as a situation. What are the infons of this situation? A conceptual graph is constructed out of concepts, references, and relations. All parts have a specific role in the description of information. For instance, the concepts describe the type of the objects. Following this idea, we propose to translate each item of a graph (concept, reference, relation) into a specific infon. We have to be careful to conserve the information as given by the graph. For instance, that we conserve the information about which referent belongs to which concept type in a graph. First we define the *conceptual graph infon language*.

Definition 4.29 Let C be a finite set of concepts and T a finite set of referents. Further, let Rel be a finite set of conceptual relations and Prm be a set of parameters. Let $\mathcal{T}$ be the union of C and T. The *conceptual graph infon language* $\mathcal{I}_{CG}(\mathcal{T})$ is defined to be the smallest set such that

(i) if $r \in Rel, \dot{\mathrm{p}}, \dot{\mathrm{q}} \in Prm$ then $\langle\langle r,\dot{\mathrm{p}},\dot{\mathrm{q}};\ 1\rangle\rangle \in \mathcal{I}_{CG}(\mathcal{T})$,
(ii) if $\mathrm{t} \in T, \dot{\mathrm{p}} \in Prm$ then $\langle\langle \mathit{Ref},\mathrm{t},\dot{\mathrm{p}};\ 1\rangle\rangle \in \mathcal{I}_{CG}(\mathcal{T})$,
(iii) if $\mathrm{C} \in C, \dot{\mathrm{p}} \in Prm$ then $\langle\langle \mathit{Type},\mathrm{C},\dot{\mathrm{p}};\ 1\rangle\rangle \in \mathcal{I}_{CG}(\mathcal{T})$.

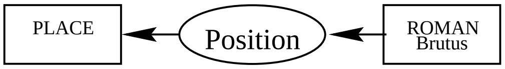
Infons as defined at item (i) are called *relation infons*. Those defined at item (ii) are called *referent infons*. Finally the infons defined at item (iii) are called *concept infons*. The conceptual graph infon language contains a set of positive infons based on a set of concepts, relations, referents, and parameters. The set of situations is defined to be the set $\mathcal{S}(\mathcal{I}_{CG}(\mathcal{T}))$ (or $\mathcal{S}_{CG}$ for short).

Given two conceptual graphs g and $h \in \mathcal{G}$, and let $\mathcal{S}_{CG}$ be the set of situations, the translation function $map : \mathcal{G} \rightarrow \mathcal{S}_{CG}$ is defined as follows:

- For each concept node u with a concept type U without a referent, the function $map(\mathrm{U})$ has as result $\{\langle\langle \mathit{Type},\mathrm{U},\dot{\mathrm{p}};\ 1\rangle\rangle\}$ with $\dot{\mathrm{p}}$ as a unique parameter and U the concept type of u.
- For each concept node u with a concept type U and a referent t, the function $map(\mathrm{U} : \mathrm{t})$ has as result $\{\langle\langle \mathit{Type},\mathrm{U},\dot{\mathrm{p}};\ 1\rangle\rangle, \langle\langle \mathit{Ref},\mathrm{t},\dot{\mathrm{p}};\ 1\rangle\rangle\}$ with $\dot{\mathrm{p}}$ a unique parameter and t the referent of concept u.
- If R is a binary relation between two conceptual graphs g and h (in that particular order) with $\langle\langle \mathit{Type},\mathrm{C},\dot{\mathrm{p}};\ 1\rangle\rangle \in map(g)$ and $\langle\langle \mathit{Type},\mathrm{D},\dot{\mathrm{q}};\ 1\rangle\rangle \in map(h)$ then

$map(gRh) = \{\langle\langle R, \dot{\mathrm{p}}, \dot{\mathrm{q}};\ 1\rangle\rangle\} \cup map(g) \cup map(h).$

Example 4.4



The translation of the above conceptual graph using *map* results in the situation $\{\langle\langle \mathit{Type}, \text{PLACE}, \dot{p};\ 1\rangle\rangle, \langle\langle \mathit{Position}, \dot{p}, \dot{q};\ 1\rangle\rangle, \langle\langle \mathit{Ref}, \text{Brutus}, \dot{q};\ 1\rangle\rangle, \langle\langle \mathit{Type}, \text{ROMAN}, \dot{q};\ 1\rangle\rangle\}$.

If we have two infons that are sharing the same parameter then we call this *corresponding* infons. For instance, the corresponding infons $\langle\langle \mathit{Type}, \text{ROMAN}, \dot{\mathrm{p}};\ 1\rangle\rangle$ and $\langle\langle \mathit{Ref}, \text{Brutus}, \dot{\mathrm{p}};\ 1\rangle\rangle\}$.

As mentioned in Chapter 3, we define two situations S, T with parameters equivalent if we can obtain two identical situations by renaming the parameters. For example, with this definition, we have that $\{\langle\langle \mathit{Position}, \dot{\mathrm{p}}, \dot{\mathrm{q}};\ 1\rangle\rangle\} \equiv \{\langle\langle \mathit{Position}, \dot{\mathrm{r}}, \dot{\mathrm{s}};\ 1\rangle\rangle\}$ and $\{\langle\langle \mathit{Position}, \dot{\mathrm{p}}, \dot{\mathrm{q}};\ 1\rangle\rangle\} \not\equiv \{\langle\langle \mathit{Position}, \dot{\mathrm{r}}, \dot{\mathrm{r}};\ 1\rangle\rangle\}$.

It is important to note here that we do not have the following property: if $S \equiv T$ then $S \cup T \equiv S$. Only if S is identical to situation T without renaming the parameters this will be the case.

The function *map* is injective: for every conceptual graph there is an unique situation. However the function is not surjective: there are situations S for which there is no g such that $map(g) = S$. Therefore we define the notion of a *graph situation*.

Definition 4.30 (Graph Situation) A situation $S \in \mathcal{S}_{CG}$ is called a *graph situation* if and only if it satisfies the following conditions for its elements:

(i) For each parameter used in the relation infons in S there exists a corresponding concept infon in S.

(ii) For each concept infon in S there exists at most one corresponding referent infon in S.

(iii) If there is more than one concept infon in S then for each concept infon there exists a corresponding relation infon in S.

(iv) For each referent infon there exists a corresponding concept infon.

(v) Each relation infon in S has exactly two parameters.

(vi) Each pair of relation infons r_i and r_j in S has a parameter in common, or there exist a list of pairs $(r_i, r_k) \ldots (r_l, r_j)$ such that $r_n \in S$ and each pair has a parameter in common.

The first condition states that each relation node should be connected with concepts. The second condition expresses that a concept type of a concept node has at most one referent. The third condition requires in case there is more than one concept node that concept nodes are connected to a relation node. The fourth item states that each referent is connected to a concept type. The fifth item limits the conceptual graph to have only dyadic relations. Finally, the last item requires that the graph is connected.

Corollary 4.5 If S and T are graph situations. Then,

(i) situation $S \cup T \cup \{\langle\langle r,\dot{\mathrm{p}},\dot{\mathrm{q}};\ 1\rangle\rangle\}$ is a graph situation if and only if there is a concept infon $\langle\langle \mathit{Type},\mathrm{C},\dot{\mathrm{p}};\ 1\rangle\rangle \in S$ and a concept infon $\langle\langle \mathit{Type},\mathrm{D},\dot{\mathrm{q}};\ 1\rangle\rangle \in T$, or there is a concept infon $\langle\langle \mathit{Type},\mathrm{C},\dot{\mathrm{q}};\ 1\rangle\rangle \in S$ and a concept infon $\langle\langle \mathit{Type},\mathrm{D},\dot{\mathrm{p}};\ 1\rangle\rangle \in T$.

(ii) situation $S \cup \{\langle\langle \mathit{Ref},\mathrm{k},\dot{\mathrm{p}};\ 1\rangle\rangle\}$ is a graph situation if and only if there is a concept infon $\langle\langle \mathit{Type},\mathrm{C},\dot{\mathrm{p}};\ 1\rangle\rangle \in S$ and $\langle\langle \mathit{Ref},\mathrm{t},\dot{\mathrm{p}};\ 1\rangle\rangle \notin S$ for $\mathrm{k} \neq \mathrm{t}$.

The *map* function is injective, as stated in the following lemma.

Lemma 4.3 If $map(G) \equiv map(H)$ then $G \equiv H$.

Proof We have to prove that for every situation, there is only one unique graph. Obviously $map(G)$ and $map(H)$ are graph situations, the situations contain only three kinds of infons, namely, referent infons, relation infons, and concept infons, with the conditions as given in Definition 4.30. The structure of the graph is conveyed by the infons, the direction of the edges between two concepts and a relation node is represented by the order of the occurrence of the parameters in the relation infon. Every concept infon can be directly translated into one concept of the graph, similar for the referent infons, which can be translated to referents of the concepts, according to the translation function. Therefore, every graph situation corresponds to a unique graph. □

The information containment holds between two concept infons $\varphi \rightarrow \psi$ if the concept type corresponding to φ is a subtype of the concept type corresponding to ψ. according to the concept taxonomy. For example, let Roman $\leq$ Person be defined in the taxonomy. Then $\langle\langle \mathit{Type},\textsc{Roman},\dot{\mathrm{p}};\ 1\rangle\rangle \rightarrow \langle\langle \mathit{Type},\textsc{Person},\dot{\mathrm{p}};\ 1\rangle\rangle$ for any parameter $\dot{\mathrm{p}}$.

Postulates

Next we propose the underlying aboutness proof system of Elen, denoted by CG_{ps}. First we start with a useful property of the graph situations.

Proposition 4.7 For all conceptual graphs $A, B \in \mathcal{G}$: if $map(A) \supseteq map(B)$ then $A \leq B$.

Proof In order to prove the proposition, we use Definition 4.27, which states that $A \leq B$ if there is a subgraph A' of A satisfying the three conditions given in Definition 4.27. If A' is a subgraph of A then by the definition of the function *map*: $map(A) \supseteq map(A')$.

Let $map(A) \equiv map(A') \cup C$. In the case that $map(A') \equiv map(B)$ all the conditions of the definition are satisfied, namely, all the conceptual relations in A' and B are identical, the concepts of A' and B are identical, and if a relation r links two concepts in B', then the same concepts are linked with r in A'. So, if $map(A) \supseteq map(B)$ there is indeed a subgraph A' of A satisfying the conditions of Definition 4.27, which proves the proposition. □

Definition 4.31 (Conceptual Graph Situation Aboutness) The aboutness proof system CG_{ps} is defined to be the triple $\langle\ \mathcal{L}(\mathcal{I}_{CG}(\mathcal{T}))$,{Reflexivity},{Set Equivalence,Left Monotonic Union,Cut,Union Containment}$\rangle$.

Theorem 4.11 The aboutness proof system CG_{ps} is sound. That is, for all conceptual graphs A, $B \in \mathcal{G}$ and $D \in \mathcal{D}$ such that $\chi(D) = A$: if $\vdash_{\mathrm{CG}_{ps}} map(A) \Box\!\!\leadsto map(B)$ then $\models_{\mathrm{CG}_m} D$ about B.

Proof First we show that the axiom Reflexivity and the rules Set Equivalence, Left Monotonic Union and Cut are sound. Secondly we show that the rule Union Containment is sound. This enable us to conclude that CG_{ps} is sound with respect to the model CG_m.

- The soundness of the axiom Reflexivity and the rules Set Equivalence, Left Monotonic Union and Cut follows directly from Proposition 4.7. In this proposition the premise $map(A) \supseteq map(B)$ allows us to conclude that $A \leq B$. The aboutness decision $map(A) \supseteq map(B)$ is defined by the sound SC_{ps} aboutness proof system. So, given the postulates of SC_{ps} we can derive that $map(A) \supseteq map(B)$ which suffices to conclude that $A \leq B$. This proves the soundness of the axiom Reflexivity and the rules Set Equivalence, Left Monotonic Union and Cut.
- We have to prove the soundness of the Union Containment rule. Given that $\varphi \rightarrow \psi$ and $S \cup \{\psi\} \Box\!\!\leadsto T$ are sound premises, the conclusion $S \cup \{\varphi\} \Box\!\!\leadsto T$ is sound. Let $S \cup \{\psi\} \equiv map(A)$ and $T \equiv map(B)$. Then, given the sound premise $S \cup \{\psi\} \Box\!\!\leadsto T$ and that the concept type C_φ corresponding to the concept infon φ is a subtype of the concept type C_ψ corresponding to the concept infon ψ, we have that $A \leq B$. Thus there is a projection of B on A, which allows us to conclude that there is a subgraph A' full-filling the requirements of the projection relation. Replacing a concept type infon by its restriction on the left-hand side of the aboutness will lead to a situation S', which is a graph situation. The corresponding graph is identical to graph A except one concept type is replaced by its restriction. Trivially this does not violate the projection relation. This suffices to conclude that Union Containment is sound.

□

Similar to the proof of Theorem 4.7 we remark that it is possible to have a situation S about another situation T without S being a graph situation. One could view these situations as a kind of graphs-under-construction situations. Since formalising the notion of graphs-under-construction is cumbersome and does not contribute to the clarity of the proof we do not digress on it here.

Theorem 4.12 The aboutness proof system CG_{ps} is complete. That is, if the following holds for all conceptual graphs A, $B \in \mathcal{G}$ and $D \in \mathcal{D}$ such that $\chi(D) = A$: if $\models_{\mathrm{CG}_m} D$ about B then $\vdash_{\mathrm{CG}_{ps}} map(A) \Box\!\!\rightsquigarrow map(B)$.

Proof We have to show that if $A \leq B$ then $map(A) \Box\!\!\rightsquigarrow map(B)$. Assume $A \leq B$, this means that A is constructed out of B using a sequence over the four graph operators. So, for each graph operator there should be a representative deduction possibility in the aboutness proof system.

- If A is a copy of B then $map(A)$ is a situation containing the same infons as $map(B)$ but this set is possibly labelled with different parameters. In this case, according to the set equivalence relation $\equiv$, we have two equivalent situations. Starting with the Reflexivity axiom, we can infer that $map(A) \Box\!\!\rightsquigarrow map(B)$ when they correspond to equivalent sets of infons.
- If A is obtained from B by the restriction operator then two cases are possible: graph A is a restricted graph of B due to a replacement of a concept type by a subtype, or in B there exists a type without a referent and in A such a referent conform to that type is added. For the first case, the rule Union Containment allows us to deduce that a subtype infon is about a type infon. Let $map(A)$ be $S \cup \{\varphi\}$, $map(B)$ be T with $C_\varphi \rightarrow C_\psi$ or stated differently, the concept type C_φ corresponding to infon φ is a specialisation of the concept type C_ψ corresponding to infon ψ. Then
$$\frac{S \cup \{\psi\} \Box\!\!\rightsquigarrow T \quad \varphi \rightarrow \psi}{S \cup \{\varphi\} \Box\!\!\rightsquigarrow T}\ \mathsf{UC}$$
Otherwise, in case of an added referent $\mathtt{t}$ to concept type C, we know that it corresponds to $map(C) \cup \{\varphi\}$ with φ the referent infon corresponding to $\mathtt{t}$. Therefore we can use the rule Left Monotonic Union and Set Equivalence in order to deduce aboutness $map(C) \cup \{\varphi\} \Box\!\!\rightsquigarrow map(C)$.
- The simplification rule, removing a relation when two concepts are linked with two identical relations, is governed in the aboutness proof system by the set equivalence rule. Namely, $S \cup \{\langle\langle R, \dot{\mathrm{p}}, \dot{\mathrm{q}};\ 1\rangle\rangle\}$ is equivalent to $S \cup \{\langle\langle R, \dot{\mathrm{p}}, \dot{\mathrm{q}};\ 1\rangle\rangle, \langle\langle R, \dot{\mathrm{p}}, \dot{\mathrm{q}};\ 1\rangle\rangle\}$. Therefore the simplification rule is modelled in the aboutness proof system.
- Two graphs that share a common concept can be externally joined to form a new graph having this common concept. So, if A is constructed from B by a join operator then $map(A) \equiv map(BxC)$, where x is representing the join operator. As mentioned before, after the join operation BxC is always equal or larger as the original graph B. Due to the properties of the map function, $map(BxC) \supseteq map(B)$.

And $map(A) \supseteq map(B)$ can be proved using Reflexivity, Left Monotonic Union, and Set Equivalence.

□

Reflections

This aboutness proof system is based on the same set postulates as the aboutness proof system of the strict coordinate model in addition with the Union Containment rule.

Proposition 4.8 In the aboutness proof system CG_{ps} we have that:

(i) The top query of CG_{ps} is the set $\{\emptyset\}$.
(ii) The bottom query of CG_{ps} is the set $\emptyset$.
(iii) The top document of CG_{ps} is the set $\{\mathcal{I}_{CG}(\mathcal{T})\}$.
(iv) The bottom document of CG_{ps} is the set $\emptyset$.

The proof of the proposition proceeds in a similar way as for the sets of the aboutness proof system SC_{ps}.

Note that since we do not have a surjective *map* function, we can not use Proposition 4.1 to derive the top and bottom elements of CG_m.

Proposition 4.9 In the IR model CG_m we have that:

(i) The top query of CG_{ps} is the set $\emptyset$.
(ii) The bottom query of CG_{ps} is the set $\emptyset$.
(iii) The top document of CG_{ps} is the set $\emptyset$.
(iv) The bottom document of CG_{ps} is the set $\emptyset$.

Proof

(i) We have to show that there is no conceptual graph that is about every graph. In case of an empty graph, that is, a graph without concepts and relations, it is not possible to build a larger graph, since there are no concepts to join with. In case of a non-empty graph, one can remove from this graph a concept and all its relations. The result can not be a specialisation of the original graph. So for each empty and non-empty graph we proved that they can not be an element of the top query set.

(ii) The specialisation relation is reflexive. Consequently there is no graph that is never about a graph.

(iii) We show that there is no graph that is about all graphs in $\mathcal{G}$. Since the empty graph is only about the empty graph, there is no graph that is about all graphs. This proves item (iii) of Proposition 4.9

(iv) The proof of item (iv) is completely analogous to the proof of item (ii).

□

One of the goals of Elen was to create a precision-oriented system in order to provide the user not with an overdose of non-relevant information but with highly precise relevant information. The aboutness decisions of CG_{ps} are therefore strict, in the sense that there are only a few possibilities to derive aboutness. As a result, often a few documents are considered to be relevant in the Elen system. In order to deliver some more documents, which still are very likely to be relevant, we inspect some new rules. These rules can be added to the aboutness proof system CG_{ps}. This sort of rules are no longer based on the projection operators, but are defined in terms of the framework. Maybe we can consider that documents which are derived with this extended system have a lower degree of relevance than documents derived with the *original* system.

Consider for instance the union of relation infon to a situation, for some relation infons R and R':

$$S \cup T \cup \{R\} \,\Box\!\!\rightsquigarrow S \cup T \cup \{R'\}.$$

This axiom implies that two situations related by a situation with the relation infon R, is about the same two situations related by situation with another relation infon R'. This axiom is already valid if $R = R'$ (applying the reflexivity axiom). Now, the situation $\{\langle\langle \mathit{Type},\mathrm{Roman},\dot{\mathrm{p}};\ 1\rangle\rangle,\ \langle\langle \mathit{Type},\mathrm{Roman},\dot{\mathrm{q}};\ 1\rangle\rangle,\ \langle\langle \mathit{Kill},\dot{\mathrm{p}},\dot{\mathrm{q}};\ 1\rangle\rangle\}$ is about $\{\langle\langle \mathit{Type},\mathrm{Roman},\dot{\mathrm{p}};\ 1\rangle\rangle,\ \langle\langle \mathit{Type},\mathrm{Roman},\dot{\mathrm{q}};\ 1\rangle\rangle,\ \langle\langle \mathit{Murder},\dot{\mathrm{p}},\dot{\mathrm{q}};\ 1\rangle\rangle\}$. Although if we replaced the 'murder' infon with the relation infon $\langle\langle \mathit{Killed},\dot{\mathrm{p}},\dot{\mathrm{q}};\ 1\rangle\rangle$ this would also be about the same situation. Of course we should be careful by adopting this axiom for every relation. It depends fully on the context and the two relations R and R' if we could suggest such an axiom. Another suggestion is to allow a parameter $\dot{\mathrm{r}}$, as a relation infon. In this case, we model that two concepts are related but we do not know in which way. The axiom can be given as: let $\dot{\mathrm{p}}$ ($\dot{\mathrm{q}}$) be a parameter used in S (and T respectively), then

$$S \cup T \cup \{\langle\langle R,\dot{\mathrm{p}},\dot{\mathrm{q}};\ 1\rangle\rangle\} \,\Box\!\!\rightsquigarrow S \cup T \cup \{\langle\langle \dot{\mathrm{r}},\dot{\mathrm{p}},\dot{\mathrm{q}};\ 1\rangle\rangle\}.$$

this idea needs an extension of the language $\mathcal{I}_{CG}(\mathcal{T})$.

An extension of CG_{ps} could be to permit the aboutness derivation between a graph and its restricted form,

Right Monotonic Relation Union (RMRU)

$$\frac{S \,\Box\!\!\rightsquigarrow T}{S \,\Box\!\!\rightsquigarrow T \cup \{\langle\langle \mathit{Ref},\mathrm{t},\dot{\mathrm{p}};\ 1\rangle\rangle\}}$$

For instance, up till now it was not allowed to conclude that the graph representing "The Roman Brutus hates a Roman" is about the graph representing "The Roman Brutus hates the Roman Caesar" because the right graph is a specialisation of the left one (rather than the opposite). With the new Right Monotonic Relation Union we are allowed to add references on the left side in order to determine aboutness. Adopting this new plausible rule can be viewed as allowing the user to *mislabel* references. A user uses a referent in the query as an example, but maybe she is looking for more general information.

Finally we want to suggest a new postulate based on situation union. In case a user is searching for a document in which "a person is driving a car" and in which "a red car" occurs, without stating that the car in which the person is driving has to be red. Because everything in ELEN is connected (the join was the only way to build up graphs) we can not express unrelated information. Therefore the rule Context-Free Union could be useful in the system. With this rule we get all that if $S \Box\!\!\rightsquigarrow T$ and $S \Box\!\!\rightsquigarrow U$ then $S \Box\!\!\rightsquigarrow T \cup U$. Or in conceptual graph words, if a graph G is about a graph H and also about a graph I, then we conclude that graph G is about graph H or about graph I.

4.7 Summary and conclusions

In this chapter we presented the formalisation of six common IR models. The formalisation mainly concerns the notion of aboutness, and the representation of the document descriptor set and query in situations. We considered eight sets $\mathbf{1}^{d}_{\mathrm{X}}, \mathbf{0}^{d}_{\mathrm{X}}, \mathbf{1}^{q}_{\mathrm{X}}$ and $\mathbf{0}^{q}_{\mathrm{X}}$ with X the IR model or the aboutness proof system. These sets correspond to typical elements of an IR model and demonstrate a specific characteristic of the model. We summarise the aboutness proof systems of the models studied in the table on the next page.

In this chapter we showed that the framework can be used to formalise several different IR models. Furthermore we have showed that soundness and completeness theorems could be proved. Given a sound and complete aboutness proof system, interesting observations could be made. It is very important to notice that we have achieved several general axiomatic definitions of aboutness. These axiomatisations provides some interesting observations that will be worked out in the next chapter.

PS	$\mathcal{L}$	*Axioms*	*Rules*
SC_{ps}	$\mathcal{L}(\mathcal{I}_{Basic}(\mathcal{T}))$	Reflexivity	Set Equivalence Left Monotonic Union Cut
C_{ps}	$\mathcal{L}(\mathcal{I}_{Basic}(\mathcal{T}))$	Singleton Reflexivity	Set Equivalence Left Monotonic Union Symmetry Strict Composition
VC_{ps}	$\mathcal{L}(\mathcal{I}_{Basic}(\mathcal{T}))$	Singleton Reflexivity	Set Equivalence Left Monotonic Union Symmetry Strict Composition
IE_{ps}	$\mathcal{L}(\mathcal{I}_{IE}(\mathcal{T}))$	Singleton Reflexivity	Set Equivalence Left Monotonic Union Symmetry Strict Composition
$\mathrm{B1}_{ps}$	$\mathcal{L}(\mathcal{I}_{B}(\mathcal{T}))$	Reflexivity	Set Equivalence Left Monotonic Union Cut $\vee$-Right Monotonic Composition
$\mathrm{B2}_{ps}$	$\mathcal{L}_2^{Ext}(\mathcal{I}_{B}(\mathcal{T}))$	Reflexivity$_1$	Set Equivalence$_1$ Left Monotonic Union$_1$ Cut$_1$ Anti-Aboutness Rule$_1$ Context-Free Union$_2$ $\vee$-Right Monotonic Composition$_2$ Aboutness Inheritance Simple Anti-Aboutness Closed World Assumption
CG_{ps}	$\mathcal{L}(\mathcal{I}_{CG}(\mathcal{T}))$	Reflexivity	Set Equivalence Left Monotonic Union Cut Union Containment

Chapter 5

Comparing IR models through their aboutness proof systems

If you can't say it in words, then you had better not whistle it in mathematics either.

C.J. van Rijsbergen & M. Lalmas, *'An Information Calculus for Information Retrieval'*.

In this chapter we combine the insights gained in the previous chapters and apply it to our main goal to devise a technique to compare and analyse IR models. The fact that many relevant IR models can be characterised by means of (sound and complete) aboutness proof systems immediately suggests to shift the focus towards the use of these proof systems. The first application is an obvious one: using aboutness proof systems one can attempt to compare IR models *theoretically* instead of *experimentally*. The known insights from logic about ways to compare formal systems and theories can be brought to bear on comparing the relative strength of IR models. The advantage of this type of comparison is that theorems could be proved, for instance expressing that one IR model is more effective than another model. Such results would not only spare us the efforts of experimentation, but more importantly, it would allow us to sidestep the controversies surrounding the experimental process. The first section presents a theoretical comparison for IR models based on the modelling work done in Chapter 4.

After we have compared the various IR models, we analyse in Section 5.2 the properties of several aboutness proof systems separately. The reason for using the more abstract aboutness proof system instead of the underlying model lies in our intention to show the syntactic properties of the models, but also in our desire to obtain some insight into the general properties of the aboutness proof systems. Therefore we study in this section some basic properties of aboutness proof systems. In particular, an important aspect of a formal reasoning system is whether it fulfils *the Principle of Monotonicity.*

The derivation of aboutness statements can be viewed as a specific reasoning process, and therefore it is interesting to investigate whether a proof system and hence, some IR model, is monotonic or not. Furthermore, we study the consequences of monotonicity and non-monotonicity from the point of view of information retrieval. We conclude the chapter with a summary and ideas for possible extensions.

5.1 The comparison of IR models

We have seen in Chapter 4 how IR models can be related to aboutness proof systems. In terms of its corresponding proof system, the collective aboutness theorems as they can be deduced form the 'theory' of an IR model. In this section we investigate in which way aboutness proof systems can be related to each other in order to make comparative statements about IR models. More in particular, we compare IR models by comparing their associated proof systems. Through studying the inferential power of proof systems, we can present a first evaluation. To make this kind of comparison formal, we introduce some additional terminology.

First we define the notion of embedding. This notion captures the idea that aboutness derivations in some system may be simulated in another system.

Definition 5.1 An aboutness proof system $\mathcal{A}_{ps} = \langle \mathcal{L}_a, Ax_a, Rule_a \rangle$ is *embedded* in an aboutness proof system $\mathcal{B}_{ps} = \langle \mathcal{L}_b, Ax_b, Rule_b \rangle$ with respect to the aboutness decision if and only if $\mathcal{L}_a \subseteq \mathcal{L}_b$ and for all situations $S, T \in \mathcal{S} \in \mathcal{L}_a$ it holds that if $\vdash_{\mathcal{A}_{ps}} S \Box\!\!\rightsquigarrow T$ then $\vdash_{\mathcal{B}_{ps}} S \Box\!\!\rightsquigarrow T$.

Informally, an aboutness proof system $\mathcal{A}_{ps}$ is embedded in an aboutness proof system $\mathcal{B}_{ps}$ if and only if the aboutness language of $\mathcal{A}_{ps}$ is a subset of the aboutness language of $\mathcal{B}_{ps}$ and all aboutness theorems of $\mathcal{A}_{ps}$ are aboutness theorems of $\mathcal{B}_{ps}$.

In logical terms it means that the theory of $\mathcal{A}_{ps}$ is a restriction of the theory of $\mathcal{B}_{ps}$ or, alternatively, that the latter is an extension of the former. The definition has a further implication in case the rules of inference of $\mathcal{A}_{ps}$ can be simulated, as is the case in many of the systems of Chapter 4, by fixed proof schemes using the rules of inference of $\mathcal{B}_{ps}$. In this case the translation of derivations in $\mathcal{A}_{ps}$ to derivations in $\mathcal{B}_{ps}$ is fully effective, and the resulting proofs in $\mathcal{B}_{ps}$ are never longer than some fixed constant factor times the proof-length in $\mathcal{A}_{ps}$. When this is the case, we say that $\mathcal{A}_{ps}$ is *conservatively embedded* in $\mathcal{B}_{ps}$.

Definition 5.2 Two aboutness proof systems $\mathcal{A}_{ps} = \langle \mathcal{L}_a, Ax_a, Rule_a \rangle$ and $\mathcal{B}_{ps} = \langle \mathcal{L}_b, Ax_b, Rule_b \rangle$ are called *equivalent* with respect to the aboutness decision if and only if $\mathcal{A}_{ps}$ is embedded in $\mathcal{B}_{ps}$ and $\mathcal{B}_{ps}$ is embedded in $\mathcal{A}_{ps}$.

Theorem 5.1 If the same set of descriptors $\mathcal{T}$ is used in the aboutness language of aboutness proof system C_{ps} and the aboutness language of aboutness proof system VC_{ps}, then C_{ps} and VC_{ps} are conservatively equivalent.

Proof The aboutness languages of C_{ps} and VC_{ps} were both defined as $\mathcal{L}(\mathcal{I}_{Basic}(\mathcal{T}))$. Furthermore, for all situations S and T we have that $\vdash_{\mathrm{C}_{ps}} S \Box\!\!\rightsquigarrow T$ if and only if $\vdash_{\mathrm{VC}_{ps}} S \Box\!\!\rightsquigarrow T$, as proved in Lemma 4.2. This implies that the aboutness proof systems C_{ps} and VC_{ps} are equivalent. The proof that C_{ps} and VC_{ps} are conservatively equivalent follows directly from the fact that the aboutness proof systems are identical.

□

For the next few theorems we need the following observation. For two languages $\mathcal{L}_a(\mathcal{I}_a)$ and $\mathcal{L}_b(\mathcal{I}_b)$ of aboutness formulae, if $\mathcal{I}_a \subseteq \mathcal{I}_b$ then $\mathcal{L}_a \subseteq \mathcal{L}_b$. For two languages $\mathcal{I}_a(\mathcal{P}_a, Rel_a, Prm_a)$ and $\mathcal{I}_b(\mathcal{P}_b, Rel_b, Prm_b)$ of infons, if $\mathcal{P}_a \subseteq \mathcal{P}_b$ and $Rel_a \subseteq Rel_b$ then $\mathcal{I}_a \subseteq \mathcal{I}_b$.

Theorem 5.2 If the same set of descriptors $\mathcal{T}$ is used in the aboutness language of aboutness proof system C_{ps} and the aboutness language of aboutness proof system IE_{ps}, then C_{ps} is conservatively embedded in IE_{ps}.

Proof The aboutness language of C_{ps} was defined as $\mathcal{L}(\mathcal{I}_{Basic}(\mathcal{T}))$ and the aboutness language of IE_{ps} was defined as $\mathcal{L}(\mathcal{I}_{IE}(\mathcal{T}))$. By Definition 4.16 we can deduce that $\mathcal{I}_{IE}(\mathcal{T})$ is a superset of $\mathcal{I}_{Basic}(\mathcal{T})$, which is sufficient to conclude that $\mathcal{L}_{Basic} \subseteq \mathcal{L}_{IE}$. Furthermore, for all situations S and T we have that if $\vdash_{\mathrm{C}_{ps}} S \Box\!\!\rightsquigarrow T$ then $\vdash_{\mathrm{IE}_{ps}} S \Box\!\!\rightsquigarrow T$, as was proved in Theorem 4.6. This implies that aboutness proof system C_{ps} is embedded in aboutness proof system IE_{ps}. The proof that C_{ps} is conservatively embedded in IE_{ps} follows directly from the fact that C_{ps} can be simulated using a subset of the rules of IE_{ps}. The latter follows because the rules of C_{ps} are a subset of the rules of IE_{ps}.

□

Theorem 5.3 If the same set of descriptors $\mathcal{T}$ is used in the aboutness language of aboutness proof system SC_{ps} and for the aboutness language of aboutness proof system $\mathrm{B1}_{ps}$, then SC_{ps} is conservatively embedded in $\mathrm{B1}_{ps}$.

Proof The aboutness language of SC_{ps} was defined as $\mathcal{L}(\mathcal{I}_{Basic}(\mathcal{T}))$ and the aboutness language of $\mathrm{B1}_{ps}$ was defined as $\mathcal{L}(\mathcal{I}_B(\mathcal{T}))$, where $\mathcal{I}_{Basic}(\mathcal{T}) = \mathcal{I}(\mathcal{P}^+(\mathcal{T}), \emptyset, \emptyset)$ and $\mathcal{I}_B(\mathcal{T}) = \mathcal{I}(\mathcal{P}(\mathcal{T}), \{\vee\}, \emptyset)$. Hence, $\mathcal{I}_{Basic}(\mathcal{T}) \subseteq \mathcal{I}_B(\mathcal{T})$, which is sufficient to conclude that $\mathcal{L}_{Basic} \subseteq \mathcal{L}_B$. Further, whenever $\vdash_{\mathrm{SC}_{ps}} S \Box\!\!\rightsquigarrow T$ then $\vdash_{\mathrm{B1}_{ps}} S \Box\!\!\rightsquigarrow T$ which has been proved in Theorem 4.7 and 4.8. This implies that SC_{ps} is embedded in $\mathrm{B1}_{ps}$. The proof that SC_{ps} is conservatively embedded in $\mathrm{B1}_{ps}$ proceeds in a similar way as the one for the conservative embedding of C_{ps} in IE_{ps} as was proved in Theorem 5.2.

□

Besides the notions of embedding and equivalence, it is also interesting to consider the notion of *minimal aboutness proof systems.* This notion is formalised as follows:

Definition 5.3 An aboutness proof system $\mathcal{A}_{ps} = \langle \mathcal{L}, Ax_a, Rule_a \rangle$ is called minimal with respect to an aboutness language $\mathcal{L}$ if and only if there exists no equivalent aboutness proof system $\mathcal{B}_{ps} = \langle \mathcal{L}, Ax_b, Rule_b \rangle$ with $Ax_b \subset Ax_a$ or $Rule_b \subset Rule_a$.

It can be argued that for all aboutness proof systems which we presented, there is a minimal equivalent one. It is not clear that in general a minimal aboutness proof system can always be found effectively and that it has attractive properties.

In Chapter 4 we observed some similarities between the aboutness proof systems associated with different IR models. Let us classify aboutness proof systems in order to look at some classes of aboutness proof systems more systematically.

Definition 5.4 Let $\mathcal{A}_{ps} = \langle \mathcal{L}, Ax, Rule \rangle$ be an aboutness proof system. Then

(i) $\mathcal{A}_{ps}$ is called an *R-system* if and only if
- $Ax = \{\text{Reflexivity}\}$ and
- $Rule = \emptyset$;

(ii) $\mathcal{A}_{ps}$ is called an *SC-system* if and only if
- $Ax = \{\text{Reflexivity}\}$ and
- $Rule = \{\text{Set Equivalence, Left Monotonic Union, Cut}\}$;

(iii) $\mathcal{A}_{ps}$ is called a *C-system* if and only if
- $Ax = \{\text{Singleton Reflexivity}\}$ and
- $Rule = \{\text{Set Equivalence, Left Monotonic Union, Symmetry, Strict Composition}\}$.

Corollary 5.1 If $\mathcal{A}_{ps} = \langle \mathcal{L}_a, Ax_a, Rule_a \rangle$ is an *R-system* and $\mathcal{B}_{ps} = \langle \mathcal{L}_b, Ax_b, Rule_b \rangle$ is an *SC-system* with $\mathcal{L}_a \subseteq \mathcal{L}_b$, then $\mathcal{A}_{ps}$ is conservatively embedded in $\mathcal{B}_{ps}$.

The reader may have noticed that R-systems and SC-systems are not embedded in C-systems. The reason is that $\emptyset \,\Box\!\!\!\leadsto \emptyset$ is an aboutness theorem of any R-system and of any SC-system, while it is not an aboutness theorem of any C-system. To circumvent this, we define the notion of a *weak embedding.*

Definition 5.5 An aboutness proof system $\mathcal{A}_{ps} = \langle \mathcal{L}_a, Ax_a, Rule_a \rangle$ is *weakly embedded* in an aboutness proof system $\mathcal{B}_{ps} = \langle \mathcal{L}_b, Ax_b, Rule_b \rangle$ with respect to the aboutness decision if and only if $\mathcal{L}_a \subseteq \mathcal{L}_b$ and for all situations $S, T \in (\mathcal{S} \setminus \emptyset) \in \mathcal{L}_a$ it holds that if $\vdash_{\mathcal{A}_{ps}} S \,\Box\!\!\!\leadsto T$ then $\vdash_{\mathcal{B}_{ps}} S \,\Box\!\!\!\leadsto T$.

Analogous to the notions of conservative embedding and equivalent aboutness proof systems we can define the notions of conservative weak embedding and of weakly equivalent aboutness proof systems. Clearly, if an aboutness proof system $\mathcal{A}_{ps}$ is embedded in the aboutness proof system $\mathcal{B}_{ps}$ then the aboutness proof system $\mathcal{A}_{ps}$ is weakly embedded in the aboutness proof system $\mathcal{B}_{ps}$.

Corollary 5.2 If $\mathcal{A}_{ps} = \langle \mathcal{L}_a, Ax_a, Rule_a \rangle$ is an SC-system and $\mathcal{B}_{ps} = \langle \mathcal{L}_b, Ax_b, Rule_b \rangle$ is a C-system with $\mathcal{L}_a \subseteq \mathcal{L}_b$, then $\mathcal{A}_{ps}$ is conservatively weakly embedded in $\mathcal{B}_{ps}$.

Theorem 5.4 If the same set of descriptors $\mathcal{T}$ is used in the aboutness language of aboutness proof system SC_{ps} and the aboutness language of aboutness proof system C_{ps}, then SC_{ps} is weakly embedded in C_{ps}.

The aboutness proof systems inspected so far can be classified using the notion of (weak) embedding as presented above. Still there are two aboutness proof systems that seem to require a further modification of the notion, namely, $\mathrm{B2}_{ps}$ and CG_{ps}. The aboutness proof system $\mathrm{B2}_{ps}$ is not captured due to the fact that the aboutness relations $\square\!\!\leadsto_1$ and $\square\!\!\leadsto_2$ are used instead of aboutness relation $\square\!\!\leadsto$. To make a comparison possible we generalise the notion of embedding in the following way.

Definition 5.6 An aboutness proof system $\mathcal{A}_{ps} = \langle \mathcal{L}_a, Ax_a, Rule_a \rangle$ is *embedded* in an aboutness proof system $\mathcal{B}_{ps} = \langle \mathcal{L}_b, Ax_b, Rule_b \rangle$ with respect to the aboutness decision if and only if $\mathcal{L}_a \subseteq \mathcal{L}_b$ and for all situations $S, T \in \mathcal{S} \in \mathcal{L}_a$ and for all i it holds that if $\vdash_{\mathcal{A}_{ps}} S \,\square\!\!\leadsto_i T$ then there exists a j such that $\vdash_{\mathcal{B}_{ps}} S \,\square\!\!\leadsto_j T$.

The derived notions of conservative embedding and equivalence can be generalised in a similar way. In case one wants to prove that an aboutness proof system with one single aboutness relation is embedded in an aboutness proof system with several aboutness relations, one should transform the single aboutness relation from $\square\!\!\leadsto$ to $\square\!\!\leadsto_1$ in order to exploit the generalised definition.

Theorem 5.5 If the same set of descriptors $\mathcal{T}$ is used in the aboutness language of aboutness proof system SC_{ps} and the aboutness language of aboutness proof system $\mathrm{B2}_{ps}$, then SC_{ps} is conservatively embedded in $\mathrm{B2}_{ps}$.

Proof The proof is completely analogous to the proof of Theorem 5.3. □

The reader may have noticed that $\mathrm{B1}_{ps}$ is not (conservatively) equivalent with $\mathrm{B2}_{ps}$, although we showed in Chapter 4 that both proof systems are sound and complete with respect to the IR model B_m. Furthermore, the aboutness languages are identical. However, at page 97 we gave as an example of an aboutness theorem of $\mathrm{B2}_{ps}$ the following theorem: $\langle\langle \mathrm{I},\mathrm{C};\ 1\rangle\rangle \,\square\!\!\leadsto_2 \{\langle\langle \mathrm{I},\mathrm{B};\ 0\rangle\rangle$. This could never be an aboutness theorem of $\mathrm{B1}_{ps}$ as noticed at page 93.

Finally, we relate the aboutness proof system CG_{ps} to one of the other aboutness proof systems. In the aboutness language of CG_{ps} profons are not contained in situations. Therefore $\mathcal{L}_{CG}$ is neither a subset nor a superset of one of the other introduced languages. To circumvent this, we define the notion of α-*embedding*.

Definition 5.7 An aboutness proof system $\mathcal{A}_{ps} = \langle \mathcal{L}_a, Ax_a, Rule_a \rangle$ is α-*embedded* in an aboutness proof system $\mathcal{B}_{ps} = \langle \mathcal{L}_b, Ax_b, Rule_b \rangle$ with respect to the aboutness decision if and only if there exists a bijective function π such that $\pi(\mathcal{L}_a) = \mathcal{L}_{a'}$ and $\mathcal{L}_{a'} \subseteq \mathcal{L}_b$ and for all situations $S, T \in \mathcal{S} \in \mathcal{L}_{a'}$ it holds that if $\vdash_{\mathcal{A}_{ps}} S \Box\!\!\rightsquigarrow T$ then $\vdash_{\mathcal{B}_{ps}} \pi(S) \Box\!\!\rightsquigarrow \pi(T)$.

In order to embedded SC_{ps} in CG_{ps} we define a function $\pi : \mathcal{L}_{Basic} \rightarrow \mathcal{L}_{Basic'}$ that maps positive profons to referent infons as follows: for any $\langle\langle \mathrm{I}, \mathrm{t}; 1 \rangle\rangle \in \mathcal{I}_{Basic}(\mathcal{T}) \in \mathcal{S}_{Basic} \in \mathcal{L}_{Basic}$ $\pi(\langle\langle \mathrm{I}, \mathrm{t}; 1 \rangle\rangle) = \langle\langle \mathit{Ref}, \mathrm{t}, \dot{\mathrm{p}}; 1 \rangle\rangle$ for some parameter $\dot{\mathrm{p}}$. Then it follows that $\mathcal{L}_{Basic'} \subseteq \mathcal{L}_{CG}$. Adopting this point of view results in an embedding relation of SC_{ps} in CG_{ps}.

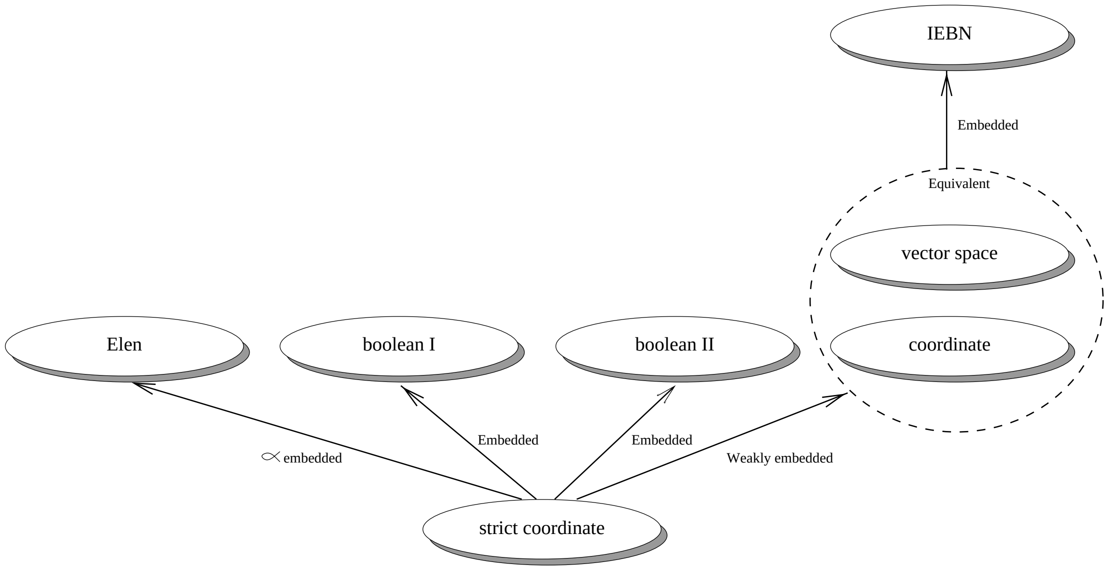
Theorem 5.6 Assume that the same set of descriptors $\mathcal{T}$ is used in aboutness language of aboutness proof system SC_{ps} and the aboutness language of aboutness proof system CG_{ps}. Furthermore, let π be a mapping function defined as above, then SC_{ps} is conservatively α-embedded in CG_{ps}.

Proof (Sketch) We refer to the proof of Proposition 4.7. There it was proved that if a graph situation S is a superset of a graph situation T, then $\vdash_{\mathrm{CG}_{ps}} S \Box\!\!\rightsquigarrow T$. □

Note that in the case the function π is defined in such a way that it maps profons to concept infons or relation infons, Theorem 5.6 is also valid.

Another point of interest is that the definition of aboutness of the vector-space model and IEBN model was strict, e.g., a document is about a query or not about a query. For example, for the vector-space model we defined query q to be about document d whenever $relcos(\chi(d), q) > 0$, but we might have required $relcos(\chi(d), q) = 1$ as another possible definition instead. If one uses the latter definition the corresponding aboutness proof system of the vector-space model would differ from VC_{ps}. Consequently, the comparison would be slightly different. We return to this subject in Section 6.2 where we investigate the situation where there are several aboutness proof systems that correspond to one IR model.

The final figure is as follows:



5.2 Analysing aboutness proof systems

For aboutness proof systems corresponding to IR models, some basic properties can be distinguished. One property, namely aboutness consistency, was already mentioned in Chapter 3. Here we want to elaborate on the following two aspects. Firstly, can we derive axioms or rules that are implicit in the aboutness proof system? These derivable axioms and rules can deliver us a deeper insight in the aboutness derivation. Secondly, in Chapter 3 an aboutness proof was viewed as a reasoning process with situation aboutness. Research in Artificial Intelligence over the past ten years has led to many new insights concerning (common-sense) reasoning processes. Being monotonic or not is an important property of formal reasoning systems. Here, we inspect aboutness proof systems on whether they fulfil the *Principle of Monotonicity*. In addition, the consequences for an aboutness proof system of being (non-)monotonic will be studied.

5.2.1 Derivable postulates

In this section we elaborate on the question whether we can derive axioms or rules that are implicit in the aboutness proof system. The axiomatisation of the IR models gives us the possibility to derive rules that offer a deeper understanding of the model under scrutiny.

Here, we consider the three classes of systems as defined in Definition 5.4. For an R-system one cannot derive interesting rules. The only observation that can be made is that for any R-system Singleton Reflexivity is a derived rule of this system. For an SC-system there are some derivable postulates that can deliver us a deeper insight in the aboutness reasoning process.

Proposition 5.1 Let $\mathcal{A}_{ps}$ be an SC-system. Then

(i) Transitivity is a derived rule of $\mathcal{A}_{ps}$,

(ii) Composition is a derived rule of $\mathcal{A}_{ps}$,

(iii) Context-Free Union is a derived rule of $\mathcal{A}_{ps}$.

Proof

(i) To prove Transitivity we have to show that, in the system $\mathcal{A}_{ps}$, $S \Box\!\!\rightsquigarrow U$ is provable from $S \Box\!\!\rightsquigarrow T$ and $T \Box\!\!\rightsquigarrow U$. So suppose that $S \Box\!\!\rightsquigarrow T$ and $T \Box\!\!\rightsquigarrow U$. Using Left Monotonic Union we deduce from $T \Box\!\!\rightsquigarrow U$ that $T \cup S \Box\!\!\rightsquigarrow U$. Since $T \cup S \equiv S \cup T$ we conclude $S \cup T \Box\!\!\rightsquigarrow U$. Finally, using Cut we conclude from $S \Box\!\!\rightsquigarrow T$ that also $S \Box\!\!\rightsquigarrow U$. Alternatively, the argument can be presented as a prooftree as follows:

$$\dfrac{S \Box\!\!\rightsquigarrow T \qquad \dfrac{\dfrac{T \Box\!\!\rightsquigarrow U}{T \cup S \Box\!\!\rightsquigarrow U}\,\textsf{LMU} \qquad T \cup S \equiv S \cup T}{S \cup T \Box\!\!\rightsquigarrow U}\,\textsf{SE}}{S \Box\!\!\rightsquigarrow U}\,\textsf{Cu}$$

(ii) To prove Composition we have to show that in the system $\mathcal{A}_{ps}$, $S \Box\!\!\rightsquigarrow T \cap U$ is provable from $S \Box\!\!\rightsquigarrow T$. Assume that $S \Box\!\!\rightsquigarrow T$. Since Reflexivity is an axiom of SC_{ps}, $T \cap U \Box\!\!\rightsquigarrow T \cap U$ is a valid premise. Left Monotonic Union allows us to deduce that $(T \cap U) \cup T \Box\!\!\rightsquigarrow T \cap U$. Then, since $(T \cap U) \cup T$ is equivalent with T, it holds that $T \Box\!\!\rightsquigarrow T \cap U$. Finally, the assumption $S \Box\!\!\rightsquigarrow T$ and the Transitivity rule are sufficient to deduce the conclusion $S \Box\!\!\rightsquigarrow T \cap U$. Alternatively, the argument can be presented as a prooftree as follows:

$$\dfrac{S \Box\!\!\rightsquigarrow T \qquad \dfrac{\dfrac{T \cap U \Box\!\!\rightsquigarrow T \cap U}{(T \cap U) \cup T \Box\!\!\rightsquigarrow T \cap U}\,\textsf{LMU} \qquad (T \cap U) \cup T \equiv T}{T \Box\!\!\rightsquigarrow T \cap U}\,\textsf{SE}}{S \Box\!\!\rightsquigarrow T \cap U}\,\textsf{Tr}$$

(iii) We have to show that in SC_{ps}, the assumptions $S \Box\!\!\rightsquigarrow T$ and $S \Box\!\!\rightsquigarrow U$ enable us to prove that $S \Box\!\!\rightsquigarrow T \cup U$. Here, we present only the prooftree:

$$\dfrac{S \Box\!\!\rightsquigarrow T \qquad \dfrac{\dfrac{S \Box\!\!\rightsquigarrow U}{S \cup T \Box\!\!\rightsquigarrow U}\,\textsf{LMU} \qquad \dfrac{\dfrac{T \cup U \Box\!\!\rightsquigarrow T \cup U}{(T \cup U) \cup S \Box\!\!\rightsquigarrow T \cup U}\,\textsf{LMU} \qquad (T \cup U) \cup S \equiv (S \cup T) \cup U}{(S \cup T) \cup U \Box\!\!\rightsquigarrow T \cup U}\,\textsf{SE}}{S \cup T \Box\!\!\rightsquigarrow T \cup U}\,\textsf{Cu}}{S \Box\!\!\rightsquigarrow T \cup U}\,\textsf{Cu}$$

□

Comparing an SC-system to a C-system, the first remark one can make is that the Cut rule does not hold in a C-system. For, given that $S \cup T$ is about U (the situation S united with T has an overlap with situation U) and S is about T (the situation S has an overlap with situation T), it does not follow that S has an overlap with U. As a counterexample, take $S = \{\phi_1, \phi_2\}$, $T = \{\phi_2, \phi_3\}$, and $U = \{\phi_3\}$ where $\phi_1 \neq \phi_2$, $\phi_1 \neq \phi_3$ and $\phi_2 \neq \phi_3$.

Transitivity is also not implied by a C-system. As a counterexample we use the previous example for the Cut rule. Given this example, one can deduce that $S \,\Box\!\!\rightsquigarrow T$ and $T \,\Box\!\!\rightsquigarrow U$ but not $S \,\Box\!\!\rightsquigarrow U$. In Chapter 3 the statement is made that Transitivity is an inherent rule of any information theoretical approach. So, in this context, using a C-system to deduce aboutness could be in conflict with desired information theoretical fundamentals.

Proposition 5.2 Let $\mathcal{A}_{ps}$ be a C-system. Then

(i) Right Monotonic Union is a derived rule of $\mathcal{A}_{ps}$,

(ii) Right Monotonic Decomposition is a derived rule of $\mathcal{A}_{ps}$.

Proof

(i) To prove Right Monotonic Union we have to show that given the assumption $S \,\Box\!\!\rightsquigarrow T$, it is provable that $S \,\Box\!\!\rightsquigarrow T \cup U$. This is easily shown by using Left Monotonic Union and Symmetry, in this order.

(ii) Given the assumption $S \,\Box\!\!\rightsquigarrow T \cap U$, with Symmetry we can deduce that $T \cap U \,\Box\!\!\rightsquigarrow S$. Left Monotonic Union allows us to deduce that $(T \cap U) \cup T$ is about S. Using Symmetry and Set Equivalence is sufficient to conclude $S \,\Box\!\!\rightsquigarrow T$.

$$\dfrac{\dfrac{\dfrac{\dfrac{S \,\Box\!\!\rightsquigarrow T \cap U}{T \cap U \,\Box\!\!\rightsquigarrow S}\,\text{Sy}}{(T \cap U) \cup T \,\Box\!\!\rightsquigarrow S}\,\text{LMU}}{S \,\Box\!\!\rightsquigarrow (T \cap U) \cup T}\,\text{Sy} \qquad (T \cap U) \cup T \equiv T}{S \,\Box\!\!\rightsquigarrow T}\,\text{SE}$$

□

The Right Monotonic Decomposition rule clearly represents the idea that in a C-system it is allowed, given that situation S is about the intersection of situation T and the situation U, to infer that situation S is about situation T. This rule is definitely not valid in any SC-system.

5.2.2 Non-monotonicity

The *non-monotonic* behaviour of rules (as encountered in *non-monotonic reasoning*) is a well-studied phenomenon in Artificial Intelligence. A rather informal definition proposed by Łukaszewicz [94] presents the idea of this sort of reasoning.

Definition 5.8 By *non-monotonic reasoning* we understand the drawing of conclusions which may be invalidated in the light of new information. A logical system is called non-monotonic iff its provability relation violates the property of monotonicity. [94]

Definition 5.9 By a non-monotonic inference pattern (a *non-monotonic rule*) we understand the following reasoning schema: "given information A, in the absence of evidence B, infer a conclusion C". [94]

An example of a non-monotonic rule in situation aboutness reasoning might be the following:

> 'Given $S \Box\!\!\rightsquigarrow T$, in the absence of the preclusion $\varphi \perp \psi$, infer $S \cup \{\varphi\} \Box\!\!\rightsquigarrow T \cup \{\psi\}$.'

The absence of $\varphi \perp \psi$ will be denoted as $\varphi \not\perp \psi$. Formally, monotonicity in terms of aboutness proof systems is defined as follows.

Definition 5.10 (Monotonicity) Let $\mathcal{A}_{ps} = \langle \mathcal{L}_n, Ax, Rule \rangle$ be an aboutness proof system.

(i) $\Box\!\!\rightsquigarrow_i$ is monotonic if and only if for all situations $S, T, U \in \mathcal{S} \in \mathcal{L}_n$:

if $\vdash_{\mathcal{A}_{ps}} S \Box\!\!\rightsquigarrow_i T$ then $\vdash_{\mathcal{A}_{ps}} S \cup U \Box\!\!\rightsquigarrow_i T$.

An aboutness relation $\Box\!\!\rightsquigarrow_i$ is called non-monotonic if and only if the aboutness relation is not monotonic.

(ii) $\mathcal{A}_{ps}$ is monotonic in its axioms if and only if for all aboutness proof systems $\mathcal{B}_{ps} = \langle \mathcal{L}_n, Ax_b, Rule \rangle$ with $Ax \subset Ax_b$ and for all situations $S, T \in \mathcal{S} \in \mathcal{L}_n$ and for all i with $1 \leq i \leq n$:

if $\vdash_{\mathcal{A}_{ps}} S \Box\!\!\rightsquigarrow_i T$ then $\vdash_{\mathcal{B}_{ps}} S \Box\!\!\rightsquigarrow_i T$.

An aboutness proof system $\mathcal{A}_{ps}$ is called non-monotonic in its axioms if and only if the aboutness proof system is not monotonic in its axioms.

(iii) $\mathcal{A}_{ps}$ is monotonic in its rules if and only if for all aboutness proof systems $\mathcal{B}_{ps} = \langle \mathcal{L}_n, Ax, Rule_b \rangle$ with $Rule \subset Rule_b$ and for all situations $S, T \in \mathcal{S} \in \mathcal{L}_n$ and for all i with $1 \leq i \leq n$:

if $\vdash_{\mathcal{A}_{ps}} S \Box\!\!\rightsquigarrow_i T$ then $\vdash_{\mathcal{B}_{ps}} S \Box\!\!\rightsquigarrow_i T$.

An aboutness proof system $\mathcal{A}_{ps}$ is called non-monotonic in its rules if and only if the aboutness proof system is not monotonic in its rules.

In case an aboutness proof system is monotonic both in its axioms and in its rules, we say that the aboutness proof system is monotonic in its postulates. Note that in case of an aboutness language $\mathcal{L}$ with only one aboutness relation one should transform the single aboutness relation from $\Box\!\!\rightsquigarrow$ to $\Box\!\!\rightsquigarrow_1$ (and $\mathcal{L}$ to $\mathcal{L}_1$) in order to use the definition of monotonicity.

An aboutness proof system of which the derivation relation is defined in the same way as classical (propositional or first-order) logic is monotonic in its postulates. However, this does not hold for arbitrary aboutness proof systems. Consider for instance an aboutness proof system that contains the Closed World Assumption rule as presented in Section 3.3.3 on page 57:

Closed World Assumption

$$\frac{S \boxtimes\!\!\rightsquigarrow_i \{\phi\} \quad \phi \bot_j \psi}{S \,\Box\!\!\rightsquigarrow_j \{\psi\}}$$

The aboutness relation $\Box\!\!\rightsquigarrow_j$ is non-monotonic and the aboutness proof system $\mathrm{B2}_{ps}$ that contains this rule is non-monotonic in its postulates. This is formally stated by the following two theorems.

Theorem 5.7 The aboutness relation $\Box\!\!\rightsquigarrow_2$ of $\mathrm{B2}_{ps}$ corresponding to the aboutness derivation of boolean retrieval, is non-monotonic. The aboutness proof system $\mathrm{B2}_{ps}$ is non-monotonic in its postulates.

Proof First we prove that $\Box\!\!\rightsquigarrow_2$ is non-monotonic. Let $S = \{\varphi\}$, $T = \{\gamma\}$, and $\psi \bot_2 \gamma$. In this case $S \,\Box\!\!\rightsquigarrow_2 T$, because $S \boxtimes\!\!\rightsquigarrow_1 \{\psi\}$ and therefore $S \,\Box\!\!\rightsquigarrow_2 \{\gamma\}$. Now consider $S' \equiv S \cup \{\psi\}$. Then $S' \,\Box\!\!\not\rightsquigarrow_2 \{\gamma\}$ because $S \,\Box\!\!\rightsquigarrow_1 \{\psi\}$. Hence the aboutness relation $\Box\!\!\rightsquigarrow_2$ of $\mathrm{B2}_{ps}$ is non-monotonic.

Next we prove that $\mathrm{B2}_{ps}$ is non-monotonic in its postulates. Let $S = \{\varphi\}$, $T = \{\gamma\}$, and $\psi \bot_2 \gamma$. In this case $S \,\Box\!\!\rightsquigarrow_2 T$. Adding the axiom $S \,\Box\!\!\rightsquigarrow_1 \{\psi\}$ to the aboutness proof system $\mathrm{B2}_{ps}$ leads to the conclusion that $S \,\Box\!\!\rightsquigarrow_2 T$ is no longer valid. Hence there is an i for which the property of monotonicity does not hold. This implies that $\mathrm{B2}_{ps}$ is non-monotonic in its postulates.

□

Note in this proof that, if we would have added the axiom $S \,\Box\!\!\rightsquigarrow_2 \{\psi\}$, then the conclusion $S \,\Box\!\!\rightsquigarrow_2 \{\varphi\}$ would still be true. Only additional information of the type $S \,\Box\!\!\rightsquigarrow_1 T$ and $\varphi \bot_2 \psi$ can withdraw aboutness conclusions.

Theorem 5.8 The aboutness proof systems SC_{ps}, C_{ps}, VC_{ps}, IE_{ps}, $\mathrm{B1}_{ps}$, and CG_{ps} are monotonic in their postulates. Furthermore, the aboutness relations of these systems are monotonic.

Proof First we prove that the aboutness relation of these systems is monotonic. All these aboutness proof systems are using an aboutness language with one aboutness relation. Furthermore they all satisfy the postulate of Left Monotonic Union. For situations S, T

and U, Left Monotonic Union supposes that from $S \square\!\!\leadsto T$ it can be concluded that $S \cup U \square\!\!\leadsto T$, which is similar to monotonicity of the aboutness relation. Therefore every aboutness relation that satisfies the postulate of Left Monotonic Union is monotonic. This implies that for each of these systems their aboutness relation is monotonic. In order to prove that the aboutness proof systems are all monotonic in their postulates we refer to the fact that they are all defined in the way the derivation relation of classical logic is defined. This implies that these systems are monotonic in their postulates. □

5.2.3 The recall of monotonic IR models

One of the goals of this thesis is a qualitative comparison of aboutness relations associated with information retrieval. An interesting possibility offered by the framework of aboutness proof systems is a qualitative assessment of quantitative retrieval measures such as recall and precision. Here we investigate, for a specific IR model, the relation between the property of monotonicity of the underlying aboutness proof system and of the aboutness relation and the quality of the recall and precision values of this model.

When one has completely characterised the aboutness relation of a specific information retrieval model by means of an aboutness proof system, one can try to prove statements such as *'addition or omission of this rule would affect recall or precision positively or negatively'.* In order to make statements about an IR model based on its underlying aboutness proof system we require some specific property of the map-function. Corresponding to the idea that extending the representation of the document should lead to new infons (or at least not to fewer infons) contained in the corresponding situation, the notion of *monotonicity* can be used.

Definition 5.11 Let x be a descriptor set. Then the function map is called *monotonic* if for all extensions x' of the descriptor set x: $map(x) \subseteq map(x')$. A function map is called non-monotonic if and only if it is not monotonic.

The definition of 'an extension of the descriptor set' depends on the representation of the document (or query) in the underlying IR model. For instance, in coordinate retrieval x' is an extension of the descriptor set x if $x \subset x'$. For the conceptual graph model, a graph that corresponds to a document (or query) can be extended by joining conceptual graphs to the original conceptual graph.

A typical example of a non-monotonic map-function is the following. Let x be a set of descriptors with $x \in \mathcal{T}$. The map-function defined as: $map(x) = \{\langle\langle \mathrm{I},\mathrm{t};\ 1\rangle\rangle \mid t \in \mathcal{T} \backslash x\}$ is non-monotonic. Extending the set x will lead to a decrease of elements of $map(x)$.

Proposition 5.3 The map-functions of SC_{ps},C_{ps},VC_{ps},IE_{ps},$\mathrm{B2}_{ps}$, and CG_{ps} and the map_2-function of $\mathrm{B1}_{ps}$ are all monotonic. The map_1-function of $\mathrm{B1}_{ps}$ is non-monotonic.

Proof We need to check for each map-function that extending the representation leads to an identical or increased number of infons of the corresponding situation. The monotonic map-functions are easy to check. We only show that the map_1-function of $\mathrm{B1}_{ps}$ is a non-monotonic map-function. The map_1-function is defined as follows:

$$map_1(x) = \{\langle\langle \mathrm{I,t;\ 1}\rangle\rangle \mid \mathrm{t} \in x \text{ and } \mathrm{t} \in \mathcal{T}\} \cup \{\langle\langle \mathrm{I,t;\ 0}\rangle\rangle \mid \mathrm{t} \notin x \text{ and } \mathrm{t} \in \mathcal{T}\}.$$

Let x be a descriptor set with $t_1 \in x$ and $t_2 \notin x$ and $\mathcal{T} = \{t_1, t_2\}$. Then $map_1(x) = \{\langle\langle \mathrm{I},t_1;\ 1\rangle\rangle, \langle\langle \mathrm{I},t_2;\ 0\rangle\rangle\}$. The extension of x with t_2 (which gives x') results in $map_1(x') = \{\langle\langle \mathrm{I},t_1;\ 1\rangle\rangle, \langle\langle \mathrm{I},t_2;\ 1\rangle\rangle\}$. Trivially, $map_1(x) \not\subseteq map_1(x')$, which suffices to conclude non-monotonicity. □

As a first start to a completely inductive theory of IR models, we give two theorems concerning the consequences for the recall value of a change of the IR model.

Theorem 5.9 If an IR model is completely described by a aboutness proof system $\mathcal{A}_{ps}$ and its aboutness relation and the used map-function(s) are monotonic, then extending the representation of the documents with more descriptors will never decrease the recall of the model.

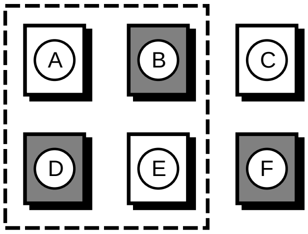
Proof On page 13, recall was defined as: $\frac{|Rel_{\mathrm{user}} \cap Ret_{\mathrm{system}}|}{|Rel_{\mathrm{user}}|}$. In this definition the set Rel_{user} is user-dependent but it is a fixed set of documents which the user judges to be relevant with respect to her information need. Let us assume that $\mid Rel_{\mathrm{user}} \mid= x$, which implies that the user indicates that there are x relevant documents in the collection. The set Ret_{system} for a aboutness proof system $\mathcal{A}_{ps}$ can be defined as $answer_q(\mathcal{A}_{ps}) = \{d \in \mathcal{D} \mid \vdash_{\mathcal{A}_{ps}} map_1(\chi(d)) \Box\!\!\leadsto map_2(q)\}$. For short, we denote $map_1(\chi(d))$ and $map_2(q)$ with S_d and S_q respectively. Let us assume that $\mid Rel_{\mathrm{user}} \cap Ret_{\mathrm{system}} \mid= y$. This implies that the aboutness proof system considers y documents about the query in harmony with the user's decision. Extending the representation of the documents with more descriptors leads in case of a monotonic function map_1 to a situation S'_d for d with $S_d \subseteq S'_d$. As $\Box\!\!\leadsto$ is assumed to be monotonic, $\vdash_{\mathcal{A}_{ps}} S_d \Box\!\!\leadsto S_q$ implies that $\vdash_{\mathcal{A}_{ps}} S'_d \Box\!\!\leadsto S_q$. However, if $\vdash_{\mathcal{A}_{ps}} S_d \Box\!\!\not\leadsto S_q$ then it could be possible that $\vdash_{\mathcal{A}_{ps}} S'_d \Box\!\!\leadsto S_q$ as well. If d is about q according to the user and the model decides that $S_d \Box\!\!\not\leadsto S_q$ and with an extended representation of d it decides that $S_d \Box\!\!\leadsto S_q$, then y will increase. Before the extension the recall was $\frac{y}{x}$. After the extension the recall becomes $\frac{z}{x}$ with $z \geq y$. This implies that for any query q the recall of the model does not decrease. □

Theorem 5.10 If an IR model is completely described by a aboutness proof system $\mathcal{A}_{ps}$ that is monotonic in its postulates, then every IR model which is completely described by an aboutness proof system obtained from $\mathcal{A}_{ps}$ by extending it with additional postulates will have a recall value that is at least as high.

Proof Similar to the proof of Theorem 5.9. By the definition of an aboutness proof system being monotonic in its postulates, extending the aboutness proof system with more postulates cannot lead to fewer aboutness theorems. In case the aboutness theorems are the same, the recall will be the same. Otherwise, if the set of aboutness theorems is extended, then if a new theorem $S_d \Box\!\!\!\rightsquigarrow S_q$ leads to a new document d in the set $answer_q(\mathcal{A}_{ps})$ and in case this new retrieved document is relevant according to the user, the recall will increase, otherwise the recall remains the same. This proves the theorem. □

The intuitive idea behind Theorem 5.10 is that in monotonic aboutness proof systems the set of aboutness theorems can only expand with the addition of axioms c.q. rules. For, due to the monotonicity, all previously derived aboutness theorems are still derivable. In other words, for any IR model as described in Theorem 5.10 the addition of new postulates can only lead to a richer theory.

Example 5.1 Consider the figure below, originally introduced in Chapter 2.



Assume that a monotonic aboutness proof system has derived that the situations A,B,D and E are about a given situation Q. By adding new postulates to the aboutness proof system it could be possible that also situation F is returned as being about the situation Q. Because the previous situations will still be returned (by the monotonicity of the aboutness relation), the additional returning of the situation F will result in a higher recall value, i.e., the recall increases from $\frac{2}{3}$ to 1.

A closer look at the example makes it also clear that one cannot make strict statements about the effect for the precision values as defined in Chapter 2. The addition of postulates to a monotonic aboutness proof system corresponding to some IR model $\mathcal{A}_m$ could increase the precision of aboutness relations, but this is only the case if $Ret_{\mathcal{A}_m} \subseteq Ret_{\mathcal{B}_m}$ and $IP_{\mathcal{B}_m,\mathcal{A}_m} > Prec_{\mathcal{A}_m}$ with $\mathcal{B}_m$ denoting the extended version of IR model $\mathcal{A}_m$, as explained in Section 2.1.1. For example, adding new postulates to aboutness proof system $\mathcal{A}_{ps}$ corresponding to IR model $\mathcal{A}_m$ could result in the determination of aboutness of situation F; this will increase the precision value to $\frac{3}{5}$, but it is also possible that only situation C would be returned as being about situation Q. In the latter case the precision value would drop to $\frac{2}{5}$.

If two models are based on the same representation language, as in the case with strict coordinate retrieval and coordinate retrieval as presented in Section 4.1 and Section 4.2 respectively, then one can relate the recall values in the case of weakly embedded aboutness proof systems as in Theorem 5.10. For instance the following theorem holds:

Theorem 5.11 If IR models $\mathcal{A}_m$ and $\mathcal{B}_m$ are completely described by aboutness proof systems $\mathcal{A}_{ps}$ and $\mathcal{B}_{ps}$ respectively, $\mathcal{A}_m$ and $\mathcal{B}_m$ are using the same indexing function χ, and aboutness proof system $\mathcal{A}_{ps}$ is weakly embedded in aboutness proof system $\mathcal{B}_{ps}$, then the recall value of model $\mathcal{A}_m$ will be less than or equal to the recall value of model $\mathcal{B}_m$.

Proof By definition of weakly embeddedness, if $\vdash_{\mathcal{A}_{ps}} S \Box\!\!\rightsquigarrow T$ then $\vdash_{\mathcal{B}_{ps}} S \Box\!\!\rightsquigarrow T$ for non-empty S and T. This implies that if $d \in answer(\mathcal{A}_{ps}, q, \mathcal{D})$ then $d \in answer(\mathcal{B}_{ps}, q, \mathcal{D})$ and trivially, if $d \in Rel_{\text{user}}$ for IR model $\mathcal{A}_m$ then $d \in Rel_{\text{user}}$ for IR model $\mathcal{B}_m$. In case $\not\vdash_{\mathcal{A}_{ps}} S \Box\!\!\rightsquigarrow T$ then it could be possible that $\vdash_{\mathcal{B}_{ps}} S \Box\!\!\rightsquigarrow T$. The rest of the proof is completely similar to the proof of Theorem 5.10.

□

Note that in Theorem 5.11 the implicit assumption is made that the query q with $map(q) = \emptyset$ is not considered for recall evaluation. The answer to the question 'what are relevant documents in case nothing is asked' is an arbitrary and subjective one. If one would answer 'all documents' then the recall of strict coordinate retrieval will be 1 for an empty query. The recall of coordinate retrieval will be 0 (as $\emptyset$ is the bottom-query of the aboutness proof system C_{ps}). So, in this particular case, Theorem 5.11 does not apply.

Corollary 5.3 The recall value of strict coordinate retrieval is always less than or equal to the recall value of coordinate retrieval.

Theorem 5.12 If IR models $\mathcal{A}_m$ and $\mathcal{B}_m$ are completely described by aboutness proof systems $\mathcal{A}_{ps}$ and $\mathcal{B}_{ps}$ respectively, $\mathcal{A}_m$ and $\mathcal{B}_m$ are using the same indexing function χ, and aboutness proof systems $\mathcal{A}_{ps}$ and $\mathcal{B}_{ps}$ are equivalent, then the respective recall and precision values of model $\mathcal{A}_m$ will be identical to the recall and precision values of model $\mathcal{B}_m$.

Proof By definition of equivalent aboutness proof systems it follows that if $\vdash_{\mathcal{A}_{ps}} S \Box\!\!\rightsquigarrow T$ then $\vdash_{\mathcal{B}_{ps}} S \Box\!\!\rightsquigarrow T$ and vice versa. Consequently, if $d \in answer(\mathcal{A}_{ps}, q, \mathcal{D})$ then $d \in answer(\mathcal{B}_{ps}, q, \mathcal{D})$ and vice versa. Thus, the recall and precision values of IR models $\mathcal{A}_m$ and $\mathcal{B}_m$ will be identical. This proves the theorem.

□

Corollary 5.4 The respective recall and precision values of coordinate retrieval and vector-space retrieval are identical.

Using the above theorems one has a first tool to make qualitative statements about the recall values of the various systems which we have studied. In the way we presented it, monotonicity thus seems to be a desirable property. Having monotonicity one can make qualitative statements, without it one cannot. Nevertheless, some authors argued that IR models should display a non-monotonic character [27, 28, 76]. In the following section we discus how non-monotonic aboutness in reality is.

5.2.4 How non-monotonic is aboutness?

In this section we concentrate on the following question: 'is aboutness monotonic?' and if so, in which way can one formalise it using the framework we have developed. We show that in information retrieval the notion of aboutness and it is present in the user's mind is typically non-monotonic. In order to handle non-monotonicity in the models of Chapter 4, some new rules will be proposed. These 'non-monotonic' rules should replace the Left Monotonic Union rule. Let us first explain why we feel that aboutness *should* be non-monotonic.

The user formulates a query, which is based on her expectation of what it returns. This expectation can be considered as a set of defaults. For instance, a user who wishes to be informed about "what is on television tonight" can formulate a query "programs". In this case, the user assumes by default that there are no other sorts of programs. If the system returns a document with the descriptor "computer programs", the user would probably reject this document.

Using Definition 5.9, one can view the use of defaults in terms of a non-monotonic aboutness derivation as follows: 'given the information that A, in the absence of evidence against default D, infer a conclusion B'. The Closed World Assumption is a rule implicitly using defaults. The default is then 'if a document is not represented by a descriptor t it is represented by the negation of t'. This is a crude approach of the system to span the users' defaults. Due to the cognitive character of the users' defaults, it is hard to formalise them in general. However, if we can formalise the users' implicit defaults the precision will probably increase. Using defaults will bring the set Ret_{system} closer to the set Rel_{user}. Thus, the fact that "programs" is about "programs" should not automatically allow us to conclude that "computer programs" is about "programs". Or, stated differently, we observed that aboutness is non-monotonic.

Next we will introduce some non-monotonic rules in order to handle the non-monotonicity of aboutness. Let us start with a formalisation of a rule that deals with a simple interpretation of a user's default. The rule is called Rational Compositional Monotonicity and states that information composition may occur only when no preclusion relationships are violated.

Rational Compositional Monotonicity (RCM)

$$\frac{S \,\Box\!\!\leadsto T \quad S \equiv \{\varphi_1, \ldots, \varphi_n\} \quad \varphi_1 \not\perp \psi \quad \ldots \quad \varphi_n \not\perp \psi}{S \cup \{\psi\} \,\Box\!\!\leadsto T}$$

Replacing the Left Monotonic Union rule by Rational Compositional Monotonicity in a SC-system or a C-system, in addition with some preclusion relations, results in a non-monotonic aboutness relation.

Let $\varphi \perp \psi$ be an axiom of a SC-system in which Left Monotonic Union is replaced by Rational Compositional Monotonicity. We can conclude in this system that $\{\varphi\} \,\Box\!\!\leadsto \{\varphi\}$ using Reflexivity. Here the conclusion that $\{\varphi\} \cup \{\psi\} \,\Box\!\!\leadsto \{\varphi\}$ does not hold.

So, given a particular information need, certain preclusion relations can be given as user's defaults. This first approach to model the non-monotonic behaviour of aboutness seems to succeed. However, given that $\varphi \perp \psi$ holds one should avoid the undesirable conclusion in which a situation occurs containing the elements φ and ψ. Such a conclusion can be achieved as follows. Starting from the Reflexivity axiom:

$$\dfrac{\dfrac{\{\omega\} \,\Box\!\!\leadsto \{\omega\}}{\{\varphi, \omega\} \,\Box\!\!\leadsto \{\omega\}}\; RCM \text{ and } SE}{\{\varphi, \psi, \omega\} \,\Box\!\!\leadsto \{\omega\}}\; RCM \text{ and } SE$$

The underlying idea that made this deduction possible is that some preclusion relations are missing. We suggest that the following preclusion rule should be adopted:

Composition Preclusion (CP)

$$\frac{S \cup \{\varphi\} \,\Box\!\!\leadsto \{\psi\} \quad \psi \perp \omega}{\varphi \perp \omega}$$

Now, if we want to have the conclusion $\{\psi, \omega\} \,\Box\!\!\leadsto \{\omega\}$ we are not allowed to conclude $\{\varphi, \psi, \omega\} \,\Box\!\!\leadsto \{\omega\}$, as for now $\varphi \perp \omega$ is an added axiom.

Let us illustrate the rules with an example. Consider a user who wishes to learn about "programs". Assume that she typically wants to be informed about aspects such as "movies", "talk shows", etc. Within this specific information need, the user would seemingly not want to be informed about "computer programs". It seems that the profon ⟨⟨Programs⟩⟩ precludes the profon ⟨⟨Computer⟩⟩ given this particular information need. Given the rule Rational Compositional Monotonicity we can conclude that {⟨⟨Programs⟩⟩, ⟨⟨Television⟩⟩} $\Box\!\!\leadsto$ {⟨⟨Programs⟩⟩} and it is then not possible to conclude that {⟨⟨Programs⟩⟩,⟨⟨Computers⟩⟩} $\Box\!\!\leadsto$ {⟨⟨Programs⟩⟩}. In case {⟨⟨Television⟩⟩,⟨⟨Programs⟩⟩} $\Box\!\!\leadsto$ {⟨⟨Television⟩⟩} is concluded, the preclusion ⟨⟨Television⟩⟩$\perp$⟨⟨Computers⟩⟩ is adopted by means of the Composition Preclusion rule. Due to this preclusion relation it is not possible to conclude then that {⟨⟨Computers⟩⟩,⟨⟨Television⟩⟩,⟨⟨Programs⟩⟩} $\Box\!\!\leadsto$ {⟨⟨Television⟩⟩}.

We have assumed that the profon $\langle\langle \mathrm{Television} \rangle\rangle$ precludes the profon $\langle\langle \mathrm{Computer} \rangle\rangle$ given the user's particular information need. The last clause 'given the user's particular information need' plays an important role in the preclusion information. Defaults based on an information need are very hard to obtain. They are time-, person-, and place-dependent and often based on non-logical grounds. Statistical information cannot be used. For example, somebody enters the query "Prime Minister". Typically the user wants to be informed about the Prime Minister of the country where she lives. Probably the user also wants to be informed about the current Prime Minister and not the one of 1980. These are the kind of defaults the IR system should take into consideration. However, if she is searching from the Netherlands in the Reuter newswire (a world-wide news document collection), the statistical information would certainly not tell her that we can associate Prime Minister with Wim Kok[1]. Instead, the concept of Prime Minister who is most favoured with news-attention is linked with her or his name, probably John Major, and that will be the default based on statistical grounds. What we can do is to extract default information of the user by using a so-called navigation process on the query. In this process the query is not entered by a sentence but via a navigation process in which the user build her request. We discuss this aspect in Chapter 6.

To conclude, this section has demonstrated that (non-)monotonicity should play a pivotal role in IR models. If an IR model is monotonic, recall predictions can be deduced. The statement than an IR model *should* be monotonic is weakened. In reality, aboutness shows a non-monotonic character under information composition. Therefore, IR systems must be conservative with regard to information composition. This conservatism should be guided by the user's defaults. The rule Left Monotonic Union should be replaced by some non-monotonic variants using the user's defaults in the derivation. The question how the user's defaults can be (automatically) derived is hard to answer. In Chapter 6 we return to this subject.

5.3 Summary and conclusions

In this chapter we showed how our framework can be applied to the fundamental analysis and comparison of IR models. There are many avenues for further research. To begin with, the investigation of further useful concepts, definitions, and theorems is needed. The definitions of embedding, non-monotonicity, and so on given in this chapter are only the beginning of a theoretical study of IR models. Detailed investigations of more definitions and theorems are needed in order to develop a complete information retrieval theory. Such a theory must ultimately enable us to accurately predict the results of possible IR models or combinations of it.

[1] Is at the moment the Prime Minister in the Netherlands.

Chapter 6

The use of the axiomatic theory for information retrieval

> *I suppose that bronzesmiths in the Bronze Age had a working knowledge of bronze, but not what we would consider a very good theoretic account of bronze. So maybe it should not surprise us to discover that the same holds for information in this Age of Information. For it does.*
>
> J. Barwise, *'The Situation in Logic'*.

In this chapter we present some ideas for an effective use of the axiomatic theory for information retrieval as laid out in the previous chapters. We investigate three directions for it: (1) a combination of IR systems based on their aboutness proof systems [70, 72], (2) a method to obtain an ordering of relevant documents based on the axiomatic definition of aboutness [68], and (3) a presentation of the use of the framework in order to model a hypermedia approach [15]. These three directions will be presented in some detail in the following sections.

In Section 6.1 we return to the notion of combining aboutness proof systems as presented in Chapter 3. The analysis of the underlying aboutness proof systems, as presented in Chapter 5, can be very useful in order to propose workable combinations. In Section 6.2 an ordering method for aboutness proof systems is presented. The ordering method provides a technique for the relative ordering of documents based on logical reasoning rather than on statistical information. In Section 6.3 we use our framework to formalise a possible integration of information retrieval and hypermedia. With the formalisation one can study some properties of the behaviour of an integrated information retrieval hypermedia system. We conclude this chapter with a summary and ideas for possible extensions.

6.1 Combining aboutness proof systems

In Chapter 1 we presented the current information retrieval paradigm and its impact on the development of new IR systems. These IR systems have the following features:

(i) there are several document-bases;

(ii) each document-base contains different types of information (for instance a 'handbook of logic' should not be treated by the system in the same manner as the 'proceedings of a conference on modal logic');

(iii) there are various types of users and there are vast differences between their information needs (for instance, in a university library there is a huge difference between a first-year student searching for relevant information and a professor searching for relevant information);

(iv) there are various kinds of search-tasks, or stated differently, there are several ways in which a user can be satisfied with the returned information (for instance, somebody may want to be informed about a subject in general or in detail).

In Chapter 3 we mentioned the basic concepts of a theory of information retrieval agents with the capability of reasoning about aboutness, based on the intuitive ideas exposed in the thesis of Van Linder [90]. The theory of agents seems tailor-made for helping to model the information retrieval problem, covering the features mentioned above. Since rational agents have the ability to reason, communicate, gather and maintain information they could be used as autonomous IR systems, operating in several document-bases. For different types of information, users and search-tasks, one could define different types of information retrieval agents. In this section we look at the aboutness proof systems from an agent-oriented perspective based on [70, 72].

We distinguish two types of agents, the *retrievers* and the *users*. The retriever agents decide whether a document representation is about a query. One can formalise a specific retriever agent for each document-base and/or type of information. User agents are agents that have a certain information need, to be satisfied by the retrievers. Similarly, for different kinds of users and search-tasks a specific type of user agent can be given. In Figure 6.1 an agent-oriented approach for an IR model is graphically depicted.

Let us explain this figure in the context of an example. Assume a user wants to be generally informed about "Caesar". She activates a specific user agent, which triggers a suitable composition of several retriever agents based on her specific information need ("Caesar") and her search task (general information). The retriever agents can be viewed as aboutness proof systems with a single aboutness relation, for instance, the one of a vector-space model or of a boolean model. Performing this search action, the selected retriever agents transfer all their aboutness theorems with respect to the query "Caesar".

The user agent decides on the basis of the retriever agents which documents should be displayed to the user. For example, the unanimous user, as defined at page 58, only returns a document if all selected retriever agents agree on the fact that the document is relevant.

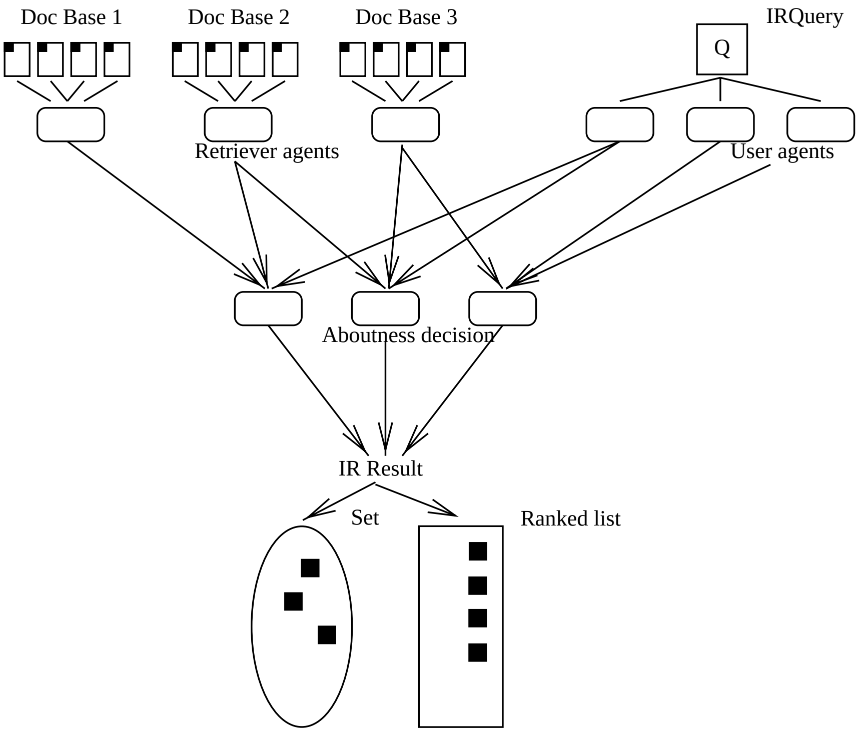


Figure 6.1: Graphical representation of an agent-oriented approach.

In Chapter 3 we defined a combined aboutness proof system $\langle \mathcal{L}_n, Ax, Rule \rangle$ with k retriever agents and l user agents with $k+l=n$. Each retriever agent represents a unique concept of aboutness. Each user agent also represents a unique concept of aboutness based on the notion of aboutness of some or all of its underlying retriever agents.

It is important to know whether two retriever agents have identical or embedded results or not. For instance, if two retriever agents always have the same aboutness theorems, a user that is relying on the two retrievers could rely on only one of them. In case one views an autonomous retriever agent as an aboutness proof system with a single aboutness relation, one could analyse the case of equivalence and embedding as proposed in Chapter 5. Two retriever agents are equivalent if and only if their corresponding aboutness proof systems are equivalent. So, it is important to know whether the retriever agents (or their corresponding aboutness proof systems) are embedded or not. For if this is the case, combining them in the way we presented in Section 3.3.4 does not yield new retrieval results. Consider the user agents as presented at page 57.

Theorem 6.1 Given two retriever agents r_1 and r_2 with their corresponding aboutness proof systems $\mathcal{A}_{ps}$ and $\mathcal{B}_{ps}$ respectively. Assume that for the aboutness proof systems $\mathcal{A}_{ps}$

and $\mathcal{B}_{ps}$ the language $\mathcal{L}$ is identical. Furthermore assume that in the combined aboutness proof system $\mathcal{C}_{ps}$ the aboutness decisions of r_1 and r_2 are combined together in a user agent u. Then

(i) if u is a *typical user* and $\mathcal{A}_{ps}$ is embedded in $\mathcal{B}_{ps}$, then u and r_2 are equivalent in terms of their aboutness decisions.

(ii) if u is a *unanimous user* and $\mathcal{A}_{ps}$ is embedded in $\mathcal{B}_{ps}$, then u and r_1 are equivalent in terms of their aboutness decisions.

Proof Note that by the definition of embedding every aboutness theorem of $\mathcal{A}_{ps}$ is also an aboutness theorem of $\mathcal{B}_{ps}$.

(i) The typical user is based on the following rule: if Φ is an aboutness theorem of one of the retriever agents, then Φ is an aboutness theorem of this user. We prove here that the aboutness theorems of $\mathcal{B}_{ps}$ are the same as the aboutness theorems of $\mathcal{C}_{ps}$. Given that Φ is an aboutness theorem of $\mathcal{C}_{ps}$, Φ should be an aboutness theorem of $\mathcal{A}_{ps}$ or of $\mathcal{B}_{ps}$. If Φ is an aboutness theorem of $\mathcal{A}_{ps}$ then by the assumption, Φ is an aboutness theorem of $\mathcal{B}_{ps}$. This is sufficient to conclude that all aboutness theorems of $\mathcal{C}_{ps}$ are aboutness theorems of $\mathcal{B}_{ps}$. The opposite, all aboutness theorems of $\mathcal{B}_{ps}$ are aboutness theorems of $\mathcal{C}_{ps}$, trivially holds by the definition of a typical user. Hence r_2 and u are equivalent in terms of their aboutness decisions.

(ii) The unanimous user is based on the following rule: if Φ is an aboutness theorem of all retriever agents, then Φ is an aboutness theorem of this user. We prove here that the aboutness theorems of $\mathcal{A}_{ps}$ are the same as the aboutness theorems of $\mathcal{C}_{ps}$. Given that Φ is an aboutness theorem of $\mathcal{C}_{ps}$, Φ should be an aboutness theorem of $\mathcal{A}_{ps}$ and of $\mathcal{B}_{ps}$. If Φ is an aboutness theorem of $\mathcal{A}_{ps}$ then by the assumption, Φ is an aboutness theorem of $\mathcal{B}_{ps}$. Thus Φ is an aboutness theorem of $\mathcal{A}_{ps}$ and of $\mathcal{B}_{ps}$. This is sufficient to conclude that u and r_1 are equivalent in terms of their aboutness decisions.

□

Informally, Theorem 6.1 captures the intuitive idea that it is not useful to use retriever agents whose corresponding aboutness proof systems are embedded, in the way given above. In order to present workable combinations, one has to inspect first whether two aboutness proof systems are embedded or not, otherwise nothing new is gained by combining them as retriever agents.

6.1.1 Filtering process

However, if two aboutness proof systems are embedded one can use this to define a sort of filtering process. For instance, consider the aboutness proof systems VC_{ps} and SC_{ps} presented in Chapter 4. We already deduced that SC_{ps} is weakly embedded in VC_{ps}.

Given a very large document-base $\mathcal{D}$, the set of relevant documents corresponding to a query q in SC_{ps} can be ideally presented with a filter as follows:

$$answer(\mathrm{SC}_{ps}, q, answer(\mathrm{VC}_{ps}, q, \mathcal{D})).$$

The idea is that first the vector-space model is used on the complete document-base and that all the resulting documents are then fed to the strict coordinate retriever. One of the advantages is that one can use a fast IR system to set bounds to a potentially enormous document-base first in order to search it more accurately with another IR system afterwards. To capture this we define a *filtering function* f-$answer$ which formalises the notion of filtering in the way presented above.

Definition 6.1 Let $\mathcal{A}_{ps}$ and $\mathcal{B}_{ps}$ be two aboutness proof systems, $\mathcal{D}$ a document-base, and q a query. The filtering function f-$answer$ of $\mathcal{A}_{ps}$ with respect to $\mathcal{B}_{ps}$ is defined by:

$$\text{f-}answer(\mathcal{A}_{ps}, \mathcal{B}_{ps}, q, \mathcal{D}) = answer(\mathcal{A}_{ps}, q, answer(\mathcal{B}_{ps}, q, \mathcal{D})).$$

We call $\mathcal{B}_{ps}$ the filter of f-$answer(\mathcal{A}_{ps}, \mathcal{B}_{ps}, q, \mathcal{D})$. Applying the filter function is called a *filtering process*. Thus applying the filter function f-$answer(\mathrm{SC}_{ps}, \mathrm{VC}_{ps}, q, \mathcal{D})$ results in the set of all documents that are first retrieved using the vector-space model and afterwards fenced in by the strict coordinate model.

Assume one has an aboutness proof system $\mathcal{A}_{ps}$ which is not fast enough. Someone wants to define a filter $\mathcal{B}_{ps}$ with the idea in mind that the filter process should not change the final set of relevant documents. This requirement means that f-$answer(\mathcal{A}_{ps}, \mathcal{B}_{ps}, q, \mathcal{D})$ $= answer(\mathcal{A}_{ps}, q, \mathcal{D})$ should hold. However note that there are some occasions in which it is desirable that f-$answer(\mathcal{A}_{ps}, \mathcal{B}_{ps}, q, \mathcal{D}) \neq answer(\mathcal{A}_{ps}, q, \mathcal{D})$, for instance if someone wants to use two systems which are disjoint with respect to the embedding relation.

One could also analyse aboutness proof systems in order to circumvent useless filters. For example the information that one aboutness proof system is embedded in another aboutness proof system can be used to prevent useless filtering processes. First we define the notion of a *useless filter*.

Definition 6.2 Let $\mathcal{A}_{ps}$ and $\mathcal{B}_{ps}$ be two aboutness proof systems. The filtering function f-$answer(\mathcal{A}_{ps}, \mathcal{B}_{ps}, q, \mathcal{D})$ is called *useless* if for all document-bases $\mathcal{D}$ and non-empty queries q

$$\text{f-}answer(\mathcal{A}_{ps}, \mathcal{B}_{ps}, q, \mathcal{D}) = answer(\mathcal{B}_{ps}, q, \mathcal{D}).$$

Informally, if the set of documents retrieved with aboutness proof system $\mathcal{B}_{ps}$ can not be fenced in by aboutness proof system $\mathcal{A}_{ps}$ then the filtering f-$answer(\mathcal{A}_{ps}, \mathcal{B}_{ps}, q, \mathcal{D})$ is useless. Note that in Definition 6.2 the empty query $(map(q) = \emptyset)$ is explicitly omitted from consideration.

Corollary 6.1 Let $\mathcal{A}_{ps}$ and $\mathcal{B}_{ps}$ be two aboutness proof systems, and assume that $\mathcal{B}_{ps}$ is weakly embedded in $\mathcal{A}_{ps}$. Then the function f-$answer(\mathcal{A}_{ps}, \mathcal{B}_{ps}, q, \mathcal{D})$ is useless.

Hence, it does not make sense to use the strict coordinate retrieval aboutness proof system SC_{ps} as a filter for a vector-space retrieval aboutness proof system by using the function f-$answer(\mathrm{VC}_{ps}, \mathrm{SC}_{ps}, q, \mathcal{D})$. This follows because the aboutness theorems of SC_{ps} (the filter) are a subset of the aboutness theorems of VC_{ps}. Thus no theorems are removed by using the system VC_{ps} after one has used the proof system SC_{ps}.

Besides proposing and inspecting filters based on aboutness proof systems, one can use filter processes to inspect aboutness proof systems.

Definition 6.3 Let $\mathcal{A}_{ps}$, $\mathcal{B}_{ps}$ be aboutness proof systems, $\mathcal{D}$ a document-base, and q a query. Given that for all document-bases $\mathcal{D}$ and queries q: f-$answer(\mathcal{A}_{ps}, \mathcal{B}_{ps}, q, \mathcal{D}) =$ f-$answer(\mathcal{B}_{ps}, \mathcal{A}_{ps}, q, \mathcal{D})$. Then,

- the aboutness proof systems $\mathcal{A}_{ps}$ and $\mathcal{B}_{ps}$ are said to *preclude* each other if and only if f-$answer(\mathcal{A}_{ps}, \mathcal{B}_{ps}, q, \mathcal{D}) = \emptyset$;
- the aboutness proof systems $\mathcal{A}_{ps}$ and $\mathcal{B}_{ps}$ are said to be *f-equivalent* if and only if f-$answer(\mathcal{A}_{ps}, \mathcal{B}_{ps}, q, \mathcal{D}) = answer(\mathcal{A}_{ps}, q, \mathcal{D})$;
- the aboutness proof systems $\mathcal{A}_{ps}$ and $\mathcal{B}_{ps}$ are said to *overlap* if and only if the systems do not preclude each other and are not f-equivalent.

Note that if two aboutness proof systems are equivalent then they are f-equivalent. The opposite does not hold. For, given that the aboutness language of a system $\mathcal{A}_{ps}$ is a superset of the aboutness language of $\mathcal{B}_{ps}$ and that the two systems have exactly the same aboutness theorems, it does not necessarily follow that $\mathcal{A}_{ps}$ and $\mathcal{B}_{ps}$ are equivalent ($\mathcal{B}_{ps}$ is conservatively embedded in $\mathcal{A}_{ps}$), but one can easily verify that $\mathcal{A}_{ps}$ and $\mathcal{B}_{ps}$ are f-equivalent.

Especially if two aboutness proof systems are in overlap, one could inspect whether the two systems are good combinations; in other words, whether the aboutness theorems of $\mathcal{A}_{ps}$ are limited in the correct way by aboutness proof system $\mathcal{B}_{ps}$, and vice versa.

6.1.2 Conclusion

In this section we have presented user and retriever agents and a filtering process based on aboutness proof systems. We showed how one can inspect workable combinations of aboutness proof systems without doing experiments. Instead of combining aboutness proof systems as rules, one could use them as filters. Combinations and filters of several IR models can be suggested on the basis of their aboutness proof systems instead of on the basis of unpredictable recall and precision values.

6.2 An ordering of aboutness proof systems

In this section we present a way to order aboutness proof systems so as to obtain an ordering of relevant documents. The ordering of the proof systems is based on the idea that each property of aboutness can be represented as an aboutness proof system, and that certain properties of aboutness are preferred over other properties.

In the aboutness proof systems considered up till now an aboutness derivation is always strict: either one can derive aboutness or one cannot. Therefore aboutness has no degrees in these proof systems. In information retrieval, however, the ordering of relevant document (ranking) is normally regarded as a necessary requirement. The classification 'relevant' and 'non-relevant' is strict and rigid. When presenting the probabilistic IR models in Chapter 4 we argued that an ordering of documents was needed because for the determination of relevance one must use imperfect knowledge; the query is not an exact match with the information need and the document representation is only a crude approximation of the document content. In a probabilistic information retrieval approach, the question whether a document is relevant to a given query is not answered by 'yes' or 'no' but by a value.

The problem that documents retrieved by our aboutness proof systems, e.g., all documents in the set $answer(\mathcal{A}_{ps}, q, \mathcal{D})$, do not have an order, also occurs in the logical IR models. Here, the relevance of a document d given a query q depends on the validity of the formula d $\texttt{about}$ q. The formula d $\texttt{about}$ q has no degrees other than true or false.

For this reason, Van Rijsbergen proposed the *Logical Uncertainty Principle* in 1986 [124] in order to extend the logical models with some uncertainty values:

> '*Given any two sentences x and y; a measure of the uncertainty of y* $\texttt{about}$ *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* $\texttt{about}$ *x.*'

The main idea is that if a system cannot logically deduce that a document d is about a query q, we have to add information to the data set until we can determine the aboutness between the document and the query. The strength of aboutness can be associated with the measure of uncertainty $P(d$ $\texttt{about}$ $q)$ which is based on how much information is added. Then according to Van Rijsbergen [124], d_1 is *preferred* over d_2 if and only if $P(d_1$ $\texttt{about}$ $q) > P(d_2$ $\texttt{about}$ $q)$. At first the data set mentioned in the principle was referring to the set of descriptors of the document representation. Later, Nie [107] suggested to calculate how much information is needed of the underlying knowledge-base to determine aboutness.

In this section we consider a different approach (based on the article [68]). We propose an ordering of documents not based on a uncertainty function but based on an ordering of aboutness proof systems. We note that the preference of the aboutness of certain documents over others with respect to a query is strongly dependent on the user. This is one advantage of our ordering technique. Typically, the role of the user is modelled as an undistinguished source of uncertainty in IR models that obtain an ordering of documents by using probability measures [54, 125, 128, 143] for handling the ordering of documents. The variations between different users cannot be accounted for because the user is not available in a symbolic form in the model.

Contrary to [124], we do not define a document d_1 to be preferred over a document d_2 if and only if $P(d_1 \text{ about } q) > P(d_2 \text{ about } q)$. In our approach, a preference of documents with respect to a query is based on an ordering of aboutness proof systems, which can be viewed as a *logical preference*. Pursuing this, we propose an order of aboutness proof systems in order to make the preference of properties of aboutness explicit. This proposal is based on the idea that each property of aboutness can be represented as an aboutness proof system, and some properties of aboutness are preferred over other properties. For instance, a document retrieved by an R-system could be preferred over a document that is retrieved by an SC-system, because for a particular user the aboutness property reflected by the R-system is more correct or intuitively more acceptable than the aboutness property reflected by an SC-system. An ordered output of documents is then obtained given a set of ordered aboutness proof systems. The retrieved document-set is ordered following the user's preferences on the aboutness properties.

To conclude, we will show an ordering technique for documents based on an ordering on aboutness proof systems. A technique is presented, that transforms an aboutness proof system into an ordered list of aboutness proof systems. The ordering of these systems is done on logical grounds only, i.e., without resorting to a quantitative formalism for uncertainty.

6.2.1 Ordering aboutness decisions

In this section we show a technique that can be used to obtain an ordering of aboutness proof systems given a set of (unordered) aboutness proof systems. The ordering represents a logical preference of documents. Using this technique the different levels of appropriateness of the aboutness properties are captured by setting preferences on the aboutness proof systems representing the properties.

Given a list of n aboutness proof systems one can define an ordering function $\pi : [1, \ldots, n] \rightarrow [1, \ldots, k]$ for some k with $1 \leq k \leq n$. This leads to an ordered list of aboutness proof systems, denoted as $\mathcal{O}$. The underlying idea is that π projects a list of aboutness proof systems onto another list in which the aboutness proof systems occur in order of preference. Let us study the ordering function π more closely. The case where $\pi : [1, \ldots, n] \rightarrow [1, \ldots, k]$ is a total surjective function results in the property that for all $1 \leq i \leq k$, there is a j with $1 \leq j \leq n$ such that $\pi(i) = j$. If this is not assumed, maybe some aboutness proof system is not mapped onto the ordering. This could be useful if an aboutness proof system of the list is not taken into consideration for the ordering, because the user thinks that the corresponding aboutness property of that particular system is not very useful. We will not consider this further.

The case in which the function is not injective, i.e., where $k \neq n$, could be useful in the case where a user likes two documents equally well. The user then considers both documents on the same level of relevance. So, if $\pi(i) = \pi(j)$ and $i \neq j$ the ith and jth aboutness proof system of the list $\mathcal{O}$ are equally preferred. In case π is bijective and hence a permutation, we call the preference of the aboutness proof systems in the list $\mathcal{O}$ a *simple order preference*. If π is bijective then the order $\sqsubset_\pi$ is a strict total order (or a linear order).

Definition 6.4 Let a list $[\mathcal{A}_{ps}, \ldots, \mathcal{Z}_{ps}]$ of n aboutness proof systems and an ordering function π be given, with $\mathcal{B}_{ps}$ denoting the ith element of the list and $\mathcal{C}_{ps}$ the jth element. Then $\mathcal{B}_{ps}$ is preferred over $\mathcal{C}_{ps}$ with respect to π if and only if $\pi(i) < \pi(j)$. If $\mathcal{B}_{ps}$ is preferred over $\mathcal{C}_{ps}$ with respect to π, we write $\mathcal{B}_{ps} \sqsubset_\pi \mathcal{C}_{ps}$.

For simplicity we write $\pi(\mathcal{B}_{ps}) = x$ when $\pi(i) = x$ and $\mathcal{B}_{ps}$ is the ith element of the list.

The preference relation on aboutness proof systems reflects the intuitive idea that a document derived with an aboutness proof system $\mathcal{B}_{ps}$ with $\pi(\mathcal{B}_{ps}) = 1$ should be considered as most relevant. Those documents derived with any aboutness proof system $\mathcal{C}_{ps}$ for which $\pi(\mathcal{C}_{ps}) = 2$ and which are not already derived with an aboutness proof system $\mathcal{B}_{ps}$ for which $\pi(\mathcal{B}_{ps}) = 1$ should be considered as second most relevant, etc. In this way one can use the ordering function π to construct a so-called *ranked document classification*.

Definition 6.5 Let a list $[\mathcal{A}_{ps}, \ldots, \mathcal{Z}_{ps}]$ of n aboutness proof systems, a document-base $\mathcal{D}$, a query q, and an ordering function $\pi : [1, \ldots, n] \rightarrow [1, \ldots, k]$ be given. Then a *ranked document classification* is the list of documents sets $[C_1, \ldots, C_k, C_{k+1}]$ such that

$$\begin{aligned}
C_1 &= \bigcup_{\pi(\mathcal{B}_{ps})=1} answer(\mathcal{B}_{ps}, q, \mathcal{D}) \\
C_k &= \bigcup_{\pi(\mathcal{B}_{ps})=k} answer(\mathcal{B}_{ps}, q, \mathcal{D}) \setminus \bigcup_{1 \leq j < k} C_j \\
C_{k+1} &= \mathcal{D} \setminus \bigcup_{1 \leq j \leq k} C_j
\end{aligned}$$

The ordering $\sqsubset_\pi$ on aboutness proof systems (obtained by the ordering function π) is meant to take into account the particular notion of relevance involved in the given retrieval situation and user's background. Or, phrased differently, for every retrieval situation or user there could be a different π.

Let us explain the concepts with an example. Consider a list of aboutness proof systems $[\mathcal{A}_1, \mathcal{A}_2, \mathcal{A}_3, \mathcal{A}_4]$. The following diagrams all represent some list $\mathcal{O}$, where $\mathcal{O}$ is the result of the ordering function π, and the arrow presents the preference relation $\sqsubset_\pi$.

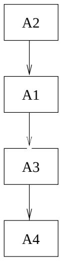


Figure 6.2: $\pi : [1, 2, 3, 4] \rightarrow [2, 1, 3, 4]$.

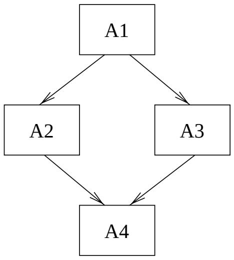


Figure 6.3: $\pi : [1, 2, 3, 4] \rightarrow [1, 2, 2, 3]$.

We have now defined a list of document sets that represents a preference of the documents of set C_i over the documents of set C_j if $i < j$. This list is constructed based on an ordering of aboutness proof systems, and so far, no numerical calculations are used. One can define the notion of *logical preference* as follows:

Definition 6.6 Let a list of n aboutness proof systems, a document-base $\mathcal{D}$, a query q, and an ordering function π be given. Then a document d_1 is *logically preferred over* a document d_2 if and only if $d_1 \in C_i$ and $d_2 \in C_j$ with $i < j$ and C_i, C_j as given in Definition 6.5. If a document d_1 is logically preferred over a document d_2 with respect to π and q we write $d_1 \prec_\pi^q d_2$.

Definition 6.7 Two ordered lists of aboutness proof systems $\mathcal{O}_1$ and $\mathcal{O}_2$ are called equivalent if and only if for all document-bases $\mathcal{D}$ and queries q, the document sets C_i of the two ranked classifications are equivalent.

Proposition 6.1 Given the list $[\mathcal{A}_1, \mathcal{A}_2, \mathcal{A}_3]$ in which $\mathcal{A}_1$ is an R-system, $\mathcal{A}_2$ is an SC-system, and $\mathcal{A}_3$ is a C-system. Then the ordering functions $\pi : [1, 2, 3] \rightarrow [3, 1, 2]$ and $\pi : [1, 2, 3] \rightarrow [3, 2, 1]$ result in equivalent ordered lists of aboutness proof systems. The same holds for the ordering functions: $\pi : [1, 2, 3] \rightarrow [2, 1, 3]$ and $\pi : [1, 2, 3] \rightarrow [2, 3, 1]$.

Proof We show the first item, leaving the second item, which is analogous to the first, to the reader. Let π be $[1, 2, 3] \rightarrow [3, 1, 2]$ then $C_1 = answer(\mathcal{A}_3, q, \mathcal{D})$. By the definition of embedding, C_2 and C_3 are empty. For if $S \,\Box\!\!\rightsquigarrow T$ is an aboutness theorem of $\mathcal{A}_1$ or $\mathcal{A}_2$ then it is an aboutness theorem of $\mathcal{A}_3$. For the function $\pi : [1, 2, 3] \rightarrow [3, 2, 1]$ the resulting list is similar, with the last two sets being empty again. □

6.2.2 Building ordered aboutness proof systems

In the previous section we proposed a technique based on aboutness proof systems that allows presenting an IR model in such a way that relevance degrees are defined in logical terms only. Moreover, the order of the aboutness proof systems could easily be obtained by asking the user for her preferences. Although the possible effects of 'an ordered aboutness proof system' have been demonstrated, we still have paid no attention to the aspects of creating an ordered aboutness proof system. We focus on the following fundamental question: 'on which set of aboutness proof systems should the π-function be based'.

We present a technique for extracting an ordering for a list of aboutness proof systems given one single aboutness proof system. Consider an aboutness proof system $\mathcal{A}_{ps}$. Without applying any method, given a query q we can already split the document-base $\mathcal{D}$ into two disjunct sets; namely, $C_1 = answer(\mathcal{A}_{ps}, q, \mathcal{D})$ and $C_2 = \mathcal{D} \setminus C_1$. In order to obtain more granularity in the classification we can construct a list of aboutness proof systems $\mathcal{O}$ with the elements of the following set:

$$\{\mathcal{B}_{ps} \mid \mathcal{B}_{ps} \text{ is (weakly) embedded in } \mathcal{A}_{ps} \ \text{ or } \ \mathcal{A}_{ps} \text{ is (weakly) embedded in } \mathcal{B}_{ps}\}.$$

In practice one should be aware of not adding aboutness proof systems to the set for which $answer(\mathcal{A}_{ps}, q, \mathcal{D}) = \emptyset$ or $answer(\mathcal{A}_{ps}, q, \mathcal{D}) = \mathcal{D}$.

For each element of the list of aboutness proof systems we have to study its corresponding aboutness property (if at all) and propose one or several ordering function(s) π based on the analysis. For specific information needs or particular types of users one can define different functions π. If the user is offered a set of ordering functions, the user is no longer a undistinguished source of uncertainty in IR models but she can play an active role in selecting an ordering of documents.

6.2.3 Logical aboutness uncertainty principle

The approach to ordering documents we proposed, uses an ordering function on a list of aboutness proof system. This leads us to propose a new logical uncertainty principle based on the principle of Van Rijsbergen. First, the following definition is needed.

Definition 6.8 Let a list $\mathcal{O}$ of n aboutness proof systems and an ordering function π be given. Furthermore let S and T be two situations. Then an aboutness proof system $\mathcal{B}_{ps}$ is called *minimal* with respect to situation S and T and list $\mathcal{O}$ if and only if $\vdash_{\mathcal{B}_{ps}} S \Box\!\!\rightsquigarrow T$ and there is no aboutness proof system $\mathcal{C}_{ps}$ in $\mathcal{O}$ such that $\vdash_{\mathcal{C}_{ps}} S \Box\!\!\rightsquigarrow T$ and $\mathcal{C}_{ps} \sqsubset_\pi \ \mathcal{B}_{ps}$.

Consider a descriptor set $\mathcal{T}$ with $A \subseteq \mathcal{T}$ and $B \subseteq \mathcal{T}$, a list of n aboutness proof systems and an ordering function π. The new principle we want to propose is this:

Logical Aboutness Uncertainty Principle:
'Given any two descriptors sets A and B and an ordered list of aboutness proof systems $\mathcal{O}$; a measure of the uncertainty of A about *B related to a given data set is determined by the minimal aboutness proof system of the list $\mathcal{O}$ we have to use to establish the truth of $map(A) \Box\!\!\rightsquigarrow map(B)$.'*

Note that we slightly redefined the principle of Van Rijsbergen. In our approach the uncertainty is determined by which aboutness proof system has to be used to determine aboutness, and not how much information has to be added. In this principle, the uncertainty measure is determined by which class a document is in, based on Definition 6.5. Of course some of the document classes could contain many documents. In order to circumvent that all the documents of one class are assumed to be equally relevant, one can use the uncertainty functions as proposed by the original logical uncertainty principle. Each class of documents, an unordered set, is ordered by means of a uncertainty function P, to have more granularity in a class. Therefore, we define a uncertainty function LP which is an extension of the function P used in the logical uncertainty principle. We have to be careful that this extension still preserves our logical ordering. Or, formally stated

Property 6.2 If $d_i \prec^q_\pi d_j$ then $LP(d_i \texttt{ about } q) > LP(d_j \texttt{ about } q)$.

We define a function LP that considers the document classification as the most important factor.

Definition 6.9 Let a list $\mathcal{O}$ of n aboutness proof systems, a document-base $\mathcal{D}$, a query q, and an ordering function π be given. Let $[C_1, \ldots, C_k]$ be the ranked document classification of k document classes based on $\mathcal{O}$, $\mathcal{D}, q$ and π. Furthermore assume a function P with a range $\langle 0, 1\rangle$ which calculates the uncertainty measure of d about q based on the logical uncertainty principle of Van Rijsbergen. The function LP is defined by:

$$LP(d \texttt{ about } q) = k \Leftrightarrow i + P(d \texttt{ about } q)$$

with $d \in C_i$.

In this definition the function P is the same for all levels, but it is worth noticing that it could be of great value to calculate different aboutness levels with different calculation functions. For example, consider the following two classes: all documents of class C_i are retrieved by an R-system, thus documents with representations that exactly matched the query. Another class C_j contains documents which are retrieved by a C-system, thus documents with representations that have some overlap with the query. Clearly, one would like to have two different functions P for the documents in classes C_i and C_j. To accommodate this, we define $LP(d \texttt{ about } q) = k \Leftrightarrow i + P_i(d \texttt{ about } q)$.

Let us explain Definition 6.9 in the context of an example.

Example 6.1 Consider the list $[\mathcal{A}_{ps}, \mathcal{B}_{ps}, \mathcal{C}_{ps}]$ in which $\mathcal{A}_{ps}$ is an R-system, $\mathcal{B}_{ps}$ is an SC-system, and $\mathcal{C}_{ps}$ is a C-system, an ordering function $\pi : [1,2,3] \rightarrow [1,2,3]$ and let map be defined as $map(x) = \{\langle\langle \mathtt{I,t;\ 1} \rangle\rangle \mid t \in x\}$. Furthermore, assume a document-base $\mathcal{D}$ containing the following documents:

doc	*doc descriptors*
d_1	$\{a, b\}$
d_2	$\{a, b, c\}$
d_3	$\{a, b, d\}$
d_4	$\{b, c\}$
d_5	$\{c, d\}$

Let q be the query $\{a, b\}$. Then $C_1 = \{d_1\}$, $C_2 = \{d_2, d_3\}$, $C_3 = \{d_4\}$, and $C_4 = \{d_5\}$ In this example $LP(d \texttt{ about } q)$ is the following:

$$\begin{aligned}
LP(d_1 \texttt{ about } q) &= 3 + P_1(d_1 \texttt{ about } q)\\
LP(d_2 \texttt{ about } q) &= 2 + P_2(d_2 \texttt{ about } q)\\
LP(d_3 \texttt{ about } q) &= 2 + P_2(d_3 \texttt{ about } q)\\
LP(d_4 \texttt{ about } q) &= 1 + P_3(d_4 \texttt{ about } q)\\
LP(d_5 \texttt{ about } q) &= P_4(d_5 \texttt{ about } q).
\end{aligned}$$

In this case d_1 is always preferred over d_2. The question whether d_2 is preferred over d_3 depends on the probability measure P_2. For instance, in case this function is using occurrence-factors it depends on whether the descriptor c occurs more often than the descriptor d. Note that in this example, given that for all $i : 0 < P(x) < 1$, the Property 6.2 holds.

6.2.4 Comparison of rankings

To show how the presented ordering technique can be used for a comparison of rankings, we briefly present the work of Wong & Yao [145, 146] as one example of how the study of ranking takes place in current information retrieval research.

In the work of Wong & Yao a user preference is defined as a binary relation $\circ\!\!>$ on $\mathcal{D}$. For $d_i, d_j \in \mathcal{D}$, $d_i \circ\!\!> d_j$ means that the user prefers d_i over d_j. They study the relation in terms of axioms, such as *asymmetry:* $d_i \circ\!> d_j \Rightarrow \neg(d_j \circ\!> d_i)$ and *negative transitivity:* $\neg(d_i \circ\!> d_j) \& \neg(d_j \circ\!> d_k) \Rightarrow \neg(d_i \circ\!> d_k)$. Furthermore, Wong & Yao study the ordering function f satisfying the following property: $d_i \circ\!> d_j \Rightarrow f(d_i) > f(d_j)$.

The result is a promising method for analysing ordering functions: for each IR model with a ranked output the ordering function can be inspected. The first-class citizen in this approach is the ordering function itself.

In our approach we are not interested in the output and the behaviour of function f, but in the *logical* sense in which $f(d_i) > f(d_j)$ is defensible for a given d_i and d_j. We are interested in the question: if a document d_i is preferred over a document d_j, e.g., $f(d_i) > f(d_j)$, is this in harmony with the[1] conception of aboutness? For instance, given a query {"Brutus", "Murder"}, an IR model that prefers a document descriptor {"Brutus", "Murder"} over a document descriptor {"Brutus", "Murder", "Caesar"} is in harmony with the assumption that a document derived with an R-system is more likely to be about the query than a document that needs an SC-system. Based on this idea, one can easily inspect the orderings of measure-based IR models, whether they are logically consistent with the 'inspectors' aboutness conception. Then one can discuss orderings based on aboutness proof systems, rather than on numerical grounds. It is also more transparent, i.e., in case of inconsistency one could point out where the ordering fails.

Example 6.2 Let us consider someone inspecting rankings of documents. She proposes that documents derived with an R-system should be preferred over documents derived with an SC-system, and that documents derived with an SC-system should be preferred over documents derived with a C-system (for the definitions see page 118). Inspecting the Index Expressions Belief Networks and the vector-space model results in the following table:

Preference relation			IEBN	VC
R-system	$\sqsubset_\pi$	C-system	Yes	Yes
R-system	$\sqsubset_\pi$	SC-system	No	Yes
SC-system	$\sqsubset_\pi$	C-system	Yes	No

This table can be interpreted as follows. VC$_{ps}$ prefers an R-system over an SC-system, but not an SC-system over a C-system. The IE$_{ps}$ prefers not an R-system over an SC-system, but an SC-system is preferred over a C-system.

This shows how it possible to compare different rankings based on our axiomatic theory.

[1] Ours, the user's, etc.

6.2.5 Theory performance

Another way to present a logical ranking of aboutness proof systems is based on Popper's procedure for theory performance. In Section 2.2.4 we described thirteen steps in order to come to a better information retrieval theory. Let us assume one starts with a certain information retrieval theory T. In an improved theory T' all the 'good' aboutness decisions of the old theory T are covered, in addition with some improvements. Sometimes there are competitive theories T'' and T''' that are both improved theories of the original theory T. If we view the theory of aboutness in terms of an aboutness proof system, we can use the inverse chronological order of the way the theory performance took place as a preference orders among theories. The order we obtain presents the belief that documents that can be considered relevant in terms of a theory T' should be preferred over documents that can be considered relevant in terms of a theory T, given that theory T' is an improved version of theory T.

6.2.6 Conclusion

In this section we have presented a technique for an ordering of suitable aboutness proof systems in order to obtain an ordered output of documents. We showed how the framework can be used for proposing or comparing a ranked output of document classification. The ordering of the documents is based on an ordering of aboutness proof systems.

6.3 A two-level hypermedia approach

In this section we give an elaborate example that shows how our theoretical framework can be applied in information retrieval. In particular, we formalise the so-called *two-level hypermedia approach*[2], which is a preliminary attempt to integrate information retrieval (aboutness decisions) and hypermedia (browsing process of the user [110].). The choice for this particular approach is based on the fact that in the two-level hypermedia approach aboutness plays a pivotal role in several different ways. The occurrence of several different aboutness decisions offers us an exquisite possibility to highlight the features of the framework presented in this thesis. We show different facets, starting with the modelling of the aboutness decisions as they occur in the paradigm. We analyse the different aboutness decisions, and show how they relate to each other. Finally, an ordering on aboutness proof systems is proposed according to the approach presented in Section 6.2. Here, the ordering is distilled automatically from the user's search actions. The basic

[2] For more detailed information we refer to [14, 23, 29].

aim of this section is to show the reader how the theoretical framework proposed in this thesis can be used in practice.

Before we start with the formalisation of the two-level hypermedia approach in terms of our framework, we discuss in Section 6.3.1 the general concepts behind the approach. In Section 6.3.2 we actually formalise the two-level hypermedia model using the theory proposed in the previous chapters. We show that in the hypermedia model several different aboutness derivations occur. Each type of derivation has its own requirements and specific properties and can be modelled as an aboutness proof system. In Section 6.3.3 and following we analyse the cooperation of the different aboutness proof systems of the hypermedia model. Finally, in Section 6.3.6 we address some further issues relating to the presented paradigm.

6.3.1 Introduction of a two-level hypermedia approach

Over the past ten years, several authors have proposed the integration of information retrieval and hypermedia [3, 29, 38, 92, 93, 141]. The *two-level hypermedia paradigm* constitutes such an integration. It consists of two levels, the *hyperbase* and the *hyperindex*. The hyperbase is a hypertext representation of the document-base $\mathcal{D}$. The hyperindex is a hypertext representation of the document representations (see Figure 6.4).

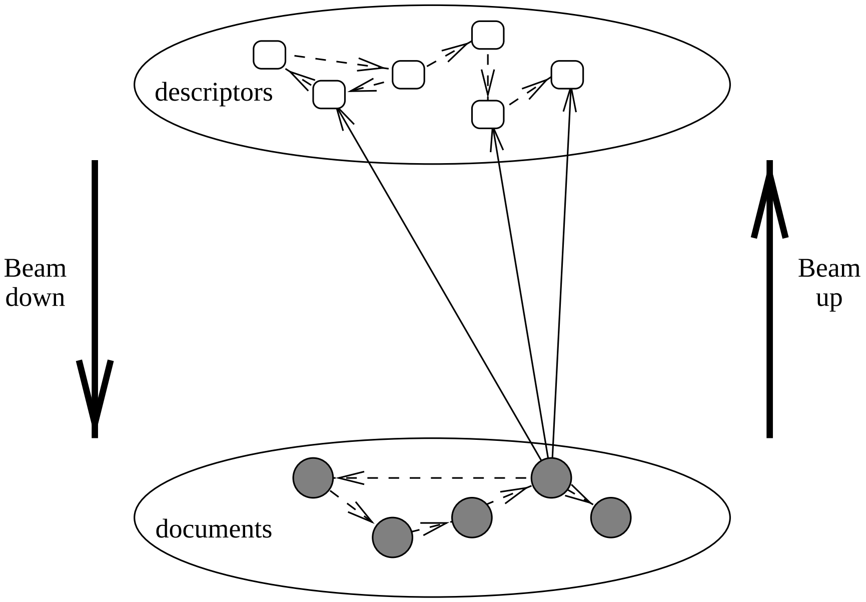


Figure 6.4: The Two-Level Hypermedia Paradigm.

A hypertext representation is a graph $\mathcal{G} = \langle N, E \rangle$, consisting of a set N of nodes and a set E of directed edges between the nodes N. In the hyperbase, the nodes are documents and the edges represent informational relations between the documents. For instance, a textual document about *"David Bowie"* could be linked with some music from one of his albums.

In Bruza's thesis [23] it is mentioned that nodes could be descriptors of the documents such as keywords or index-expressions, or the nodes could be elements of a thesaurus. Edges could then represent associative links, hierarchical links, refinements, or enlargements.

In Kheirbek & Chiaramella's work [78, 79] there are two different hyperindices, one with nodes representing types of concepts, the other with nodes representing types of conceptual relations.

Other than in a query language where the user enters a query by typing in a number of keywords, a two-level hypermedia approach allows a user to perform a search-action by travelling through the hyperindex along the edges until she is satisfied with a node. This frees the user from having to know the system's concepts (as represented by the nodes) in advance. For each step in the hyperindex, she chooses a node that is more likely to represent her information need than the current node. This process is called *query-by-navigation* (QBN) [13, 14, 16, 17, 29]. In short, QBN is the process whereby the user (as a searcher) constructs a query by travelling through the Hyperindex along the edges. The sequence of decisions taken during this travel is called a *search path*.

After finishing a search action the user wants to be informed about her constructed query. With a *beam-down* operation the user goes from the hyperindex to the hyperbase. The documents that are about a query q are accessible for the user. In a simple approach, q will be the last node of the search path. In more advanced approaches the search path of the user can be used as a context for constructing an expanded query q.

After reading, listing, or viewing some documents and following some document-links, the user possibly wants to perform another search action from the perspective of the current document. A *beam-up* operation will take the user from the hyperbase back to the hyperindex. In a simple approach the accessible node will be the one that is the representation of the current document.

A well-known phenomenon concerning the hypermedia paradigm is the so-called *feeling of getting lost in Hyperspace*. This occurs when a searcher loses track of the original information need as a result of the large amount of steps taken through the hypertext. Even though some information in the document base may be non-relevant, in many cases a user cannot resist the temptation of 'just taking a quick look'. This often leads to a departure from the concepts which were originally searched for.

The problem for the builders of a hypermedia-system is how to prevent a user from the feeling of getting lost. Based on the work of Berger [13, 14, 16, 17], one should aim to prevent the user from becoming lost in Hyperspace by examining a user's behaviour and making a statement concerning the areas of the document base in which the user *might* be interested. During the search for information, one can try to *guide* the user towards the areas in which she might be interested. The word ‘guide’ is emphasised, since under no circumstance should the retrieval system automatically place the user in the hypothetical search target. The user is allowed to make the decision in which direction the search is to be continued.

So far we presented the aspects we want to formalise in our framework. First we present the hypermedia paradigm in terms of our framework.

6.3.2 The two-level hypermedia situated paradigm

There are several approaches to use the two-level hypermedia paradigm for modelling a retrieval system, dependent on whether the focus is on the user [14, 16], on the domain knowledge [78, 79] or on the task domain [97, 98].

As mentioned in the introduction of this section, our focus will obviously be the aboutness relation as it occurs in the paradigm. One can distinguish four components:

the hyperbase: a graph $\mathrm{B} = \langle N_b, E_b \rangle$ with $N_b = \mathcal{D}$, the document-base. The edges E are links between the documents;

the hyperindex: a graph $\mathrm{I} = \langle N_i, E_i \rangle$ with $N_i = \mathcal{S}$ the set of descriptor-sets. The edges E are links between descriptor-sets;

the beam-down operator: an operator that searches for documents which are about a descriptor-set (possibly given a certain search path);

the beam-up operator: an operator that presents the document representation (set of descriptors) of a given document.

We have now presented four components, based on which we formalise the two-level hypermedia approach. Next we consider each component individually and formalise it in terms of the framework.

The hyperbase

As usual, a document carries information. Information in one document could be related to information in another document. In terms of this thesis the *relatedness* is an aboutness relation. Information in one document is about information in another document. One can connect these two documents using links. Whether two documents are *informationally* related to each other or not, is hard to determine automatically. The problems, mentioned in Chapter 3, that arise in the formalisation of relevance also arise when trying to formalise information links.

So-called multi-media authoring systems (see for instance [59]) are helping an author to create hypermedia applications, but those systems are never automatically creating information links. We define a hyperbase to be a tuple $\mathrm{B} = \langle \mathcal{D}, E_b \rangle$ with E_b a set of pairs such that if $(d_i, d_j) \in E_b$ then there is an informational link from document d_i to the document d_j. We call it the set of the document-links. For the rest of this section, we assume that the set of documents is fixed and that document-links are manually created.

The hyperindex

The hyperindex is a graph $\mathrm{I} = \langle N_i, E_i \rangle$. In contrast to the document-links of the hyperbase, the links in the hyperindex are often created automatically. As mentioned in Chapter 3, the representation of a document d consists of a set $\chi(d)$ of descriptors. In Chapter 4 we transformed each descriptor set into an abstract situation. Here we assume that χ is a function that maps each document directly to a set of infons. Given the representation functions χ_o of Chapter 4, one can easily transform the introduced representation functions (for instance the one of the index-expressions) as follows: $\chi(d) = map(\chi_o(d))$ with map the corresponding map-function. Then, given a document from the document-base, the indexing process χ directly delivers an abstract situation representing the information of the document.

The infon set $\mathcal{I}$ will be $\{\varphi \mid \varphi \in \chi(d) \text{ and } d \in \mathcal{D}\}$, and as a consequence in our aboutness language the set of abstract situations will be the powerset $\wp(\mathcal{I})$ of $\mathcal{I}$. One could consider each element of $\wp(\mathcal{I})$ (which is a set of infons, or stated differently, which is an abstract situation) to be a node of the hyperindex. In this approach abstract situations representing no information at all are also elements of the powerset and consequently a node of the graph. We limit the set of abstract situations N_i in the hyperindex to only those situations that are equivalent to or that are a subset of a representation of a document. More formally, $N_i = \{S \mid S \subseteq \chi(d) \text{ and } d \in \mathcal{D}\}$. Hence, in this approach the nodes of the hyperindex are abstract situations. We have one final remark to make about the abstract situations as to how they are used as nodes of the hyperindex. Set equivalent situations should occur in the hyperindex as one single node. These situations represent the same information. Therefore we assume the rule Set Equivalence to be implicit for all aboutness proof systems intended for hyperindices.

One can generate links between the nodes using an aboutness proof system. Given an aboutness proof system $\mathcal{A}_{ps}$, if $\vdash_{\mathcal{A}_{ps}} S \,\square\!\!\rightsquigarrow T$ then $(T, S) \in E_i$. In words, if S is about T then the user can travel from node T to S, as the information present at node T is also about the information present at node S. Note that we do not assume beforehand that also $(S, T) \in E_i$ since symmetry could be an undesirable property of an hyperindex link.

Now we can define a *corresponding hyperindex* as follows:

Definition 6.10 Given an aboutness proof system $\mathcal{A}_{ps} = \langle \mathcal{L}, Ax, Rule \rangle$, the *correspond-*

ing hyperindex I is defined to be the tuple $\langle N_i, E_i \rangle$ with $N_i = \{S \mid S \subseteq \chi(d) \text{ and } d \in \mathcal{D}\}$ and $E_i = \{(T, S) \in N_i \times N_i \mid \vdash_{\mathcal{A}_{ps}} S \Box\!\!\rightsquigarrow T\}$.

Furthermore, we call $\mathcal{A}_{ps}$ the corresponding aboutness proof system of the hyperindex I if it satisfies the requirement: if $(T, S) \in E_i$ then $\vdash_{\mathcal{A}_{ps}} S \Box\!\!\rightsquigarrow T$.

Various aboutness proof systems can be proposed for constructing hyperindex-links. Important with respect to the aboutness proof system is that a hyperindex is used for a query-by-navigation process of the user. As a consequence, aboutness decisions between nodes should be taken in small steps, as big steps easily confuse the user. For this reason the axiom $S \cup T \Box\!\!\rightsquigarrow S$ would not be appropriate. For, by adopting this axiom a user can travel in one step from node $\{\phi_1\}$ to node $\{\phi_1, \phi_2\}$, or with the same ease travel towards the node $\{\phi_1, \phi_2, \ldots, \phi_{100}\}$. Transitivity is also an undesired property of the hyperindex. Given that $(T, S) \in E_i$ and $(U, T) \in E_i$ then $(U, S) \in E_i$ would lead to the possibility of big steps and an overload of edges. Therefore we want to avoid the property of transitivity. Reflexive links allow the user to travel without leaving the node. Therefore, one should also avoid to propose links in the hyperindex that are reflexive.

Of great benefit will be the property that each link can be defined with one specific postulate. Given this property one can label each link with a particular postulate. The search path of a user can then be modelled as a sequence of logical steps of the user. We call this property *uniqueness* that is defined as follows:

Definition 6.11 Let an aboutness proof system $\mathcal{A}_{ps} = \langle \mathcal{L}, Ax, Rule \rangle$ and a corresponding hyperindex $\mathrm{I} = \langle N_i, E_i \rangle$ be given. The aboutness relation of $\mathcal{A}_{ps}$ is called *unique* if it satisfies exactly one of the following requirements:

(i) if $(T, S) \in E_i$ then $S \Box\!\!\rightsquigarrow T \in Ax$ and $Rule = \emptyset$,

(ii) if $(T, S) \in E_i$ then $S \Box\!\!\rightsquigarrow T$ is the conclusion of exactly one rule $R \in Rule$.

In words, a unique aboutness relation presents a hyperindex where each edge corresponds to a unique axiom or rule in the hyperindex logic. Then we can state about each link exactly which axiom or rule it corresponds to and which property it reflects. Note that if there is an aboutness theorem in Ax then the set $Rule$ is empty. Consequently if $S \Box\!\!\rightsquigarrow T$ is the conclusion of a rule, in the premises of this rule there are no aboutness theorems. This uniqueness property ‘protects’ us from cases in which we can derive undesirable links by a combination of axioms and rules. We define a function $corr$ which, given an aboutness proof system and two situations connected with an edge, returns a new aboutness proof system that contains the corresponding unique axiom or rule of the aboutness proof system. For instance, given a hyperindex I with an edge $(T, S) \in E$ and a corresponding aboutness proof system $\mathcal{A}_{ps} = \langle \mathcal{L}, Ax, Rule \rangle$, then $corr(\mathcal{A}_{ps}, T, S) = \langle \mathcal{L}, \{A\}, \{R\} \rangle$ if $\vdash_{\mathcal{A}_{ps}} S \Box\!\!\rightsquigarrow T$, $S \Box\!\!\rightsquigarrow T \notin Ax$ and the only way we could prove $S \Box\!\!\rightsquigarrow T$ is by applying rule R using premise A.

An aboutness proof system $\mathcal{A}_{ps}$ that is used for constructing a hyperindex and meeting all the requirements mentioned above, is called a *hyperindex logic* (HIL). The requirements of such a logic can be formalised as follows:

Definition 6.12 Given a hyperindex I and a corresponding aboutness proof system $\mathcal{A}_{ps}$. The aboutness proof system $\mathcal{A}_{ps}$ is a *hyperindex logic* if and only if the aboutness relation of $\mathcal{A}_{ps}$ is irreflexive, not transitive and unique.

Proposition 6.2 The aboutness proof systems introduced in Chapter 4 are not hyperindex logics.

Proof This follows directly from the fact that all aboutness proof systems presented in Chapter 4 have a reflexive aboutness decision.

□

Next we inspect some axioms and rules which are suitable to build a hyperindex logic. The first intuitive hypertext-link one may want to formalise is a so-called *refinement link*. By means of this link the user can travel from a node S to a node S' by extending node S to S'. Modelling the fact that a user wants to make her request more specific, the node S' contains more infons, and therefore covers more information than node S.

This link can be formalised by the Left Singleton Monotonic Union rule:

Left Singleton Monotonic Union (LSMU)

$$\frac{S \not\equiv S \cup \{\varphi\}}{S \cup \{\varphi\} \,\Box\!\!\leadsto S}$$

Note that given an aboutness proof system which contains this rule, one could conclude that $\{\langle\langle\text{Brutus}\rangle\rangle,\langle\langle\text{Caesar}\rangle\rangle\} \,\Box\!\!\leadsto \{\langle\langle\text{Brutus}\rangle\rangle\}$. The premise of the rule is needed in order to avoid reflexivity. If the premise was not required one could conclude that $\{\varphi\} \cup \{\varphi\} \,\Box\!\!\leadsto \{\varphi\}$, which results in a reflexive edge.

Using this particular rule, the conclusion $\{\varphi, \psi, \omega\} \,\Box\!\!\leadsto \{\omega\}$ is not allowed. If we want to derive the last aboutness decision one should adopt the axiom $S \cup T \,\Box\!\!\leadsto S$. However, besides reflexivity this axiom causes transitivity, which is an undesired property of the hyperindex logic.

The opposite of refinement is called *enlargement*. In this case the user wants to generalise the representation of the node where she is at the moment. This link is formalised by the Right Singleton Monotonic Union rule:

Right Singleton Monotonic Union (RSMU)

$$\frac{S \not\equiv S \cup \{\varphi\}}{S \,\Box\!\!\leadsto S \cup \{\varphi\}}$$

Using these two rules we allow the user to construct a query by travelling through the Hyperindex, meanwhile refining or enlarging the nodes.

As Bruza suggested [23] a thesaurus could be useful for creating a hyperindex. In case we want to adopt the information available in a thesaurus the following containment rule could be useful:

Left Singleton Containment (LSC)

$$\frac{\varphi\rightarrow\psi \quad S \not\equiv S \cup \{\varphi, \psi\}}{S \cup \{\varphi\} \,\Box\!\!\rightsquigarrow S \cup \{\psi\}}$$

Here, information in a thesaurus is transformed as a kind of information containment. Note, that $S \not\equiv S \cup \{\varphi, \psi\}$ is required in order to avoid the conclusion $S \,\Box\!\!\rightsquigarrow S$. Remark also that here the $\rightarrow$ is not exactly an information containment relation, as one should avoid the case wherein $\varphi\rightarrow\varphi$, because this causes reflexivity. For this reasons, we do not refer to $\rightarrow$ as the information containment relation, but as a thesaurus containment relation. A thesaurus containment relation is irreflexivity. We assume that the thesaurus information is represented as a set K of axioms.

If Left Singleton Containment is a rule of $\mathcal{A}_{ps}$ then the information that $\langle\langle\text{Brutus}\rangle\rangle \rightarrow\langle\langle\text{Roman}\rangle\rangle \in K$ results in an edge $(\{\langle\langle\text{Roman}\rangle\rangle\}, \{\langle\langle\text{Brutus}\rangle\rangle\})$ of the corresponding hyperindex. Note that the rule Union Containment, as introduced on page 51, can not be adopted. Adopting this rule one cannot travel from the node $\{\langle\langle\text{Roman}\rangle\rangle\}$ to the node $\{\langle\langle\text{Brutus}\rangle\rangle\}$ as the reflexivity axiom ($\emptyset \,\Box\!\!\rightsquigarrow \emptyset$) is not an axiom of a hyperindex logic.

The consequence of using the thesaurus containment relation in a hyperindex logic are twofold. Firstly, we require that the containment relation is irreflexive, in order to forestall reflexive links. Secondly, in case one uses a thesaurus, an update of the set N_i of nodes is needed. For instance, if $\{\langle\langle\text{Brutus}\rangle\rangle, \langle\langle\text{Caesar}\rangle\rangle\}$ is an element of N_i, then given the information that $\langle\langle\text{Brutus}\rangle\rangle\rightarrow\langle\langle\text{Roman}\rangle\rangle$ one should extend N_i with the node $\{\langle\langle\text{Roman}\rangle\rangle, \langle\langle\text{Caesar}\rangle\rangle\}$. More formally, in case a thesaurus is used, the set N_i of nodes is defined to be the set $\{S \mid S \subseteq \chi(d)$ and $d \in \mathcal{D}\} \cup \{S \mid S \equiv \{\phi_1, \ldots, \phi_k, \varphi_1, \ldots, \varphi_n\}$ and $S' \equiv \{\phi_1, \ldots, \phi_k, \varphi'_1, \ldots, \varphi'_n\}$ and for all $1 \leq i \leq n$ ($\varphi'_i\rightarrow\varphi_i$ or $\varphi_i\rightarrow\varphi'_i$) and $S' \subseteq \chi(d)$ and $d \in \mathcal{D}$ and $k \geq 0$ and $n \geq 1\}$.

Example 6.3 Given the singleton set of document descriptors $\{\{a, b\}\}$, and $a\rightarrow c$, then $N_i = \{\{a\}, \{b\}, \{c\}, \{a, b\}, \{c, b\}\}$.

In order to make it possible to travel from the node $\{\langle\langle\text{Brutus}\rangle\rangle\}$ to the node $\{\langle\langle\text{Roman}\rangle\rangle\}$ one could adopt the rule Right Singleton Containment:

Right Singleton Containment (RSC)

$$\frac{\varphi\rightarrow\psi \quad S \not\equiv S \cup \{\varphi, \psi\}}{S \cup \{\psi\} \,\Box\!\!\rightsquigarrow S \cup \{\varphi\}}$$

So far we presented four rules: Left Singleton Monotonic Union, Right Singleton Monotonic Union, Left Singleton Containment, Right Singleton Containment.

Theorem 6.2 Given K a set of axioms with elements of the type $\varphi \rightarrow \psi$ representing thesaurus information, the aboutness proof system $\mathcal{A}_{ps} = \langle \mathcal{L}, K, \{$ Left Singleton Monotonic Union, Right Singleton Monotonic Union, Left Singleton Containment, Right Singleton Containment$\}\rangle$ is a hyperindex logic.

Proof Let I $= \langle N_i, E_i \rangle$ be the corresponding hyperindex of the aboutness proof system $\mathcal{A}_{ps}$. We show successively show that $\mathcal{A}_{ps}$ meets all three requirements of a hyperindex logic.

(i) The aboutness relation of $\mathcal{A}_{ps}$ is irreflexive. We have to show that, for all $S \in N_i : (S, S) \notin E_i$, or stated differently that for all $S \in N_i : S \Box\!\!\rightsquigarrow S$ is not an aboutness theorem. Note that the axioms are not aboutness theorems. The premises of the four rules do not contain aboutness relations and since the conclusion of each rule is an aboutness theorem, we may conclude that an aboutness theorem is the result of applying one rule. Inspecting the four rules one can see that for Left Singleton Monotonic Union we have that if $S \Box\!\!\rightsquigarrow T$ then $S \supset T$ and consequently $T \neq S$. For Right Singleton Monotonic Union we have that $S \subset T$ and the conclusion that $T \neq S$ holds. For Left Singleton Containment we have that if $S \Box\!\!\rightsquigarrow T$ by using the knowledge $\varphi \rightarrow \psi$ then $\varphi \in S$ and $\psi \in T$ with $\varphi \neq \psi$ (since the thesaurus containment relation is irreflexive), the conclusion that $T \neq S$ is valid. The proof for Right Singleton Containment is completely analogous to the proof of Left Singleton Containment. This proves that the aboutness relation of $\mathcal{A}_{ps}$ is irreflexive.

(ii) To prove that the aboutness relation of $\mathcal{A}_{ps}$ is not transitive, i.e., there is a $S \Box\!\!\rightsquigarrow T$ and $T \Box\!\!\rightsquigarrow U$ and $S \Box\!\!\not\rightsquigarrow U$, we give a case in which this happens. Take for example the following three situations: $S = \{\psi_1, \psi_2, \psi_3\}$, $T = \{\psi_1, \psi_2\}$, and $U = \{\psi_1\}$. Then $S \Box\!\!\rightsquigarrow T$, $T \Box\!\!\rightsquigarrow U$ but $S \Box\!\!\not\rightsquigarrow U$.

(iii) Finally we have to prove that for each link there is a unique axiom or a unique rule of $\mathcal{A}_{ps}$, i.e., $S \Box\!\!\rightsquigarrow T$ can only be a conclusion using one specific axiom or rule. In the proof of item (i) we already concluded that an aboutness theorem is the result of applying one rule, since the axioms are not aboutness theorems. We have to show that an aboutness theorem is derived by only one rule. Inspecting the four rules one can easily verify that this is the case. This proves that the uniqueness of the aboutness relation.

□

Beam-down operator

The beam-down operator provides a link from a node of the hyperindex (an abstract situation) to a node of the hyperbase (a document). Or, in more advanced applications, the operator provides a link from one node of the hyperindex to several nodes of the hyperbase, where the latter collection of nodes could be displayed as an ordered list. In the formalisation of the hyperindex and the hyperbase the beam-down operator provides a link from an abstract situation to a document or a set of documents. We have presented in Chapter 4 several aboutness proof systems that decide whether $S \Box\!\!\rightsquigarrow T$, or in words, is situation S about situation T. This proof system was a model of an IR model that decides whether a document representation $\chi(d)$ is about a query q. Here we have to inspect whether a situation T is about a document d. We can easily use adopt the aboutness proof systems of Chapter 4 as the beam-down operator, although now one does not have to prove that $map(\chi(d)) \Box\!\!\rightsquigarrow map(q)$ because q is already a situation. Here we have to prove whether $map(\chi(d)) \Box\!\!\rightsquigarrow T$ with T a node of the hyperindex and d a node of the hyperbase. We refer to an aboutness proof system formalising the beam-down operator as a *beam-down logic* (BML).

Definition 6.13 Let $\mathrm{I} = \langle N_i, E_i \rangle$ be a hyperindex and $\mathrm{B} = \langle N_b, E_b \rangle$ be a hyperbase. Furthermore, let $\mathcal{B}_{ps}$ be a beam-down logic. Then the result of a beam-down action from node T of the hyperindex I will be a set of nodes D of the hyperbase B with D defined as:

$$\{d \in N_b \mid \vdash_{\mathcal{B}_{ps}} S \Box\!\!\rightsquigarrow T \text{ and } S = \chi(d) \text{ and } S, T \in N_i\}.$$

Note that different aboutness proof systems can be selected, resulting in different beam-down logics. The choice of a proof system could depend on the type of user, her information need, the followed search path, and so on.

Performing the beam-down operator and travelling from one node T of the hyperindex to a node d of the hyperbase is denoted by $T \bigtriangledown d$. Note that the situation T in Definition 6.13 could be the last node of the search path. However, as mentioned in the introduction of this section, one of the advantages of the query-by-navigation process is that one could also take into account the previous steps of the user's search path. We will return to this subject in Section 6.3.2.

Beam-up operator

Next we introduce the beam-up operator. This operator provides a link from a node of the hyperbase to a node of the hyperindex. For simplicity, we consider this operator to be the index function χ. In terms of the presented formalisation, given a document node

d of the hyperbase, the beam-up operator delivers a node S (an abstract situation) of the hyperindex. Here S represents the information inherent in the document d.

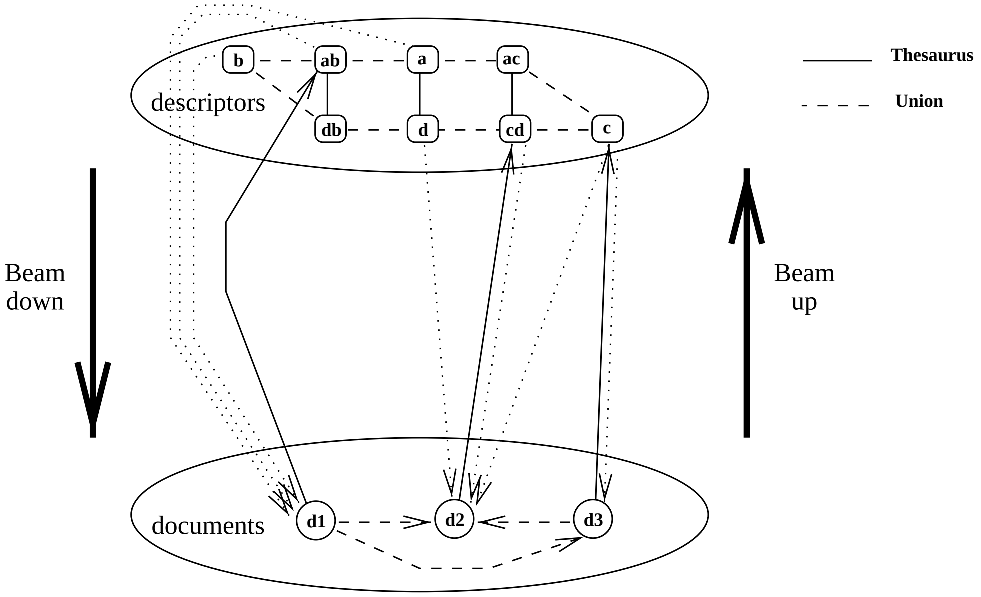
Definition 6.14 Let $\mathrm{B} = \langle N_b, E_b \rangle$ be a hyperbase and $\mathrm{I} = \langle N_i, E_i \rangle$ be a hyperindex. Furthermore, let $\chi : N_b \to N_i$ be a representation function. Then the result of a beam-up action from node d of the hyperbase B will be the node S of the hyperindex I with $S = \chi(d)$.

The beam-up action from d to S is denoted by $d \bigtriangleup S$.

Let us summarise the notions introduced thus far with an example.

Example 6.4 Let the hyperbase B consist of the graph $\langle \{d_1, d_2, d_3\}, \{(d_1, d_2), (d_1, d_3), (d_3, d_2)\}$, and let the representation function χ be such that the result of the indexing process is given by $\chi(d_1) = \{a, b\}$, $\chi(d_2) = \{c, d\}$, and $\chi(d_3) = \{c\}$.

Furthermore, let $K = \{a \to d\}$. Given the hyperindex logic $\mathcal{A}_{ps} = \langle \mathcal{L}, K, \{$ Left Singleton Monotonic Union, Right Singleton Monotonic Union, Left Singleton Containment, Right Singleton Containment$\} \rangle$ and the beam-down logic SC$_{ps}$ as presented in Chapter 4, the result is depicted in the following figure:



Note that there is a huge difference in beam-down links if we would add the rule Union Containment to the beam-down logic. Then the knowledge-base K could also be used for retrieval purposes (and not only for providing links in the hyperindex).

To conclude we have four different aboutness relations:

1. a hyperindex logic $\mathcal{A}_{ps}$, formalising the aboutness relation between the nodes of the hyperindex;

2. a given set of aboutness relations between the nodes of the hyperbase;
3. a beam-down logic $\mathcal{B}_{ps}$, formalising the aboutness relation between a node of the hyperindex and a set of nodes of the hyperbase;
4. a representation function χ, formalising the aboutness relation between a node of the hyperbase and a node of the hyperindex.

As an attentive reader may have noticed, we have introduced two logics so far, the aboutness proof systems $\mathcal{A}_{ps}$ and $\mathcal{B}_{ps}$. We propose that the search path of the user is the third logic involved in the paradigm. The formalisation of the search path will be the topic of the next section.

Search path

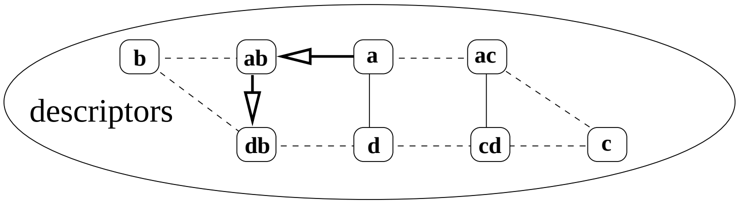
Query-by-navigation is the process whereby the searcher constructs a representation of her information need by travelling through the hyperindex along the links. A search path is a sequence of followed links by the user. We denote a search path by $S_1 \rightsquigarrow S_2 \rightsquigarrow \ldots \rightsquigarrow S_k$, which presents us the information that the user travelled from S_1 to S_2, and from S_2 to S_3 etc. Finally the user arrived at node S_k. Formally,

Definition 6.15 Given a hyperindex $\mathrm{I} = \langle N_i, E_i \rangle$. A *search path* of length k is a sequence of k linked nodes $S_1 \rightsquigarrow \ldots \rightsquigarrow S_k$ such that for all S_i, S_{i+1} with $1 \leq i$ and $i+1 \leq k$, $(S_i, S_{i+1}) \in E_i$.

In our approach, the edges of the hyperindex are formalised by postulates of a hyperindex logic. Each edge can be identified with one postulate of the hyperindex logic, and consequently each step of the search path can be labelled with one aboutness proof system. For identification we can use the function $corr(\mathcal{A}_{ps}, S_i, S_{i+1})$. Given a search path, one can transform this path into a set of aboutness proof systems. One can compose the aboutness proof systems into a single one by taking all the axioms and rules of the several aboutness proof systems together. All the axioms and rules of the aboutness proof systems occurring in the search path are members of the new aboutness proof system. An aboutness proof system that is constructed in this way given a search path is termed a *search path logic* (SPL).

Definition 6.16 Let a hyperindex $\mathrm{I} = \langle N_i, E_i \rangle$ and a corresponding hyperindex logic $\mathcal{A}_{ps} = \langle \mathcal{L}, Ax, Rule \rangle$ and a search path P of length k be given. Then a *search path logic* is the aboutness proof system $\langle \mathcal{L}, Ax', Rule' \rangle$ with $Ax' = \bigcup_{1 \leq i \leq k-1} Ax''$ and $Rule' = \bigcup_{1 \leq i \leq k-1} Rule''$ where $\langle \mathcal{L}, Ax'', Rule'' \rangle = corr(\mathcal{A}_{ps}, S_i, S_{i+1})$ for $S_i \rightsquigarrow S_{i+1} \in P$.

Example 6.5 The arrows in the figure below represent a user's search path P in the hyperindex of the previous example. The search path logic is $\langle \mathcal{L}, \{a \rightarrow d\}, \{$Left Singleton Monotonic Union, Right Singleton Containment$\}\rangle$.



One might wonder in which way the search path should influence the beam-down operator. It would be very useful if one can distill an aboutness logic $\mathcal{B}_{ps}$ from the search path logic and afterwards use this logic as input for the beam-down logic. For instance, assume a user who uses Left Singleton Monotonic Union derivation steps only. The search path logic could be used as an important factor for choosing a Right Monotonic Union oriented aboutness proof system as the beam-down logic.

Another possibility of using the search path is not to work towards the logic $\mathcal{B}_{ps}$ but towards the final query. For instance, consider the situation in which the user ends up at node S and she wants to beam-down. Using Definition 6.13 all the documents of which the representations are about S are considered to be relevant. Maybe one can add extra information to S based on the covered route through the hyperindex.

Our claim is that one can use the search path logic of a user for a better retrieval performance. What has to be done is to find a correlation between the different sorts of aboutness proof systems. Here the work presented in Chapter 5 can be used. In the next section we will go into detail about these aspects.

6.3.3 Relating aboutness proof systems

Usually it is hard to find *the* logic behind a certain search behaviour. For example, if a searcher visits a node that was visited before during the search, does this imply that she is lost in hyperspace? Or is she just sure about the information represented by this node and not about the next one? So far three different kinds of *logics* were introduced, the logic associated with the beam-down operator (the BDL), the associated logic with the way the hyperindex is created (the HIL), and the associated logic with the search path of a user (the SPL). In this section we show how these three logics are related.

If one would take the beam-down logic to be identical to the search path logic, the beam-down logic would not be very useful. For the search path logic is a subset of the hyperindex logic, and consequently, aboutness is not transitive and irreflexive. In Chapter 3 we claimed that these properties are desirable for an aboutness proof system that is used to determine aboutness between a document and a request. Therefore we need for each link in the hyperindex an aboutness proof system representation that allows us to derive documents. We present a function SP_{ps} that maps a search path logic onto an aboutness proof system. The function is defined by $\mathrm{SP}_{ps}(\langle \mathcal{L}_a, Ax_a, Rule_a \rangle) = \langle \mathcal{L}_b, Ax, Rule_b \rangle$ where for each possible search path logic there is a corresponding aboutness proof system. For instance one could suggest the following instantiations:

- If $Ax_a = \emptyset$ and $Rule_a = \{\text{Left Singleton Monotonic Union}\}$ then $Ax_b = \{\text{Reflexivity}\}$ and $Rule_b = \{\text{Left Monotonic Union}, \text{Cut}, \text{Set Equivalence}\}$.
- If $Ax_a = \emptyset$ and $Rule_a = \{\text{Left Singleton Monotonic Union}, \text{Right Singleton Monotonic Union}\}$ then $Ax_b = \{\text{Singleton Reflexivity}\}$ and $Rule_b = \{\text{Left Monotonic Union}, \text{Symmetry}, \text{Strict Composition}, \text{Set Equivalence}\}$.
- If $Ax_a = K$ and $Rule_a = \{\text{Right Singleton Monotonic Union}, \text{Left Singleton Containment}\}$ then $Ax_b = \{\text{Singleton Reflexivity}\} \cup K$ and $Rule_b = \{\text{Right Montonic Union}, \text{Left Containment}\}$.

The first item maps a search path logic of a user who has only used the links created by the rule Left Singleton Monotonic Union onto an aboutness proof system that is based on the strict-coordinate model. At the second item we propose that a search path logic of a user with Left Singleton Monotonic Union and Right Singleton Monotonic Union rules should be mapped into an aboutness proof system that is based on the coordinate model. A user did not only refine the nodes but also made some generalisation steps, therefore, the aboutness of this user could be based on an overlap. This kind of choices are subjective, and could only be proposed after an in-depth investigation of user behaviours.

As we noticed in the previous section we can also use the search path logic for an extension of the query q rather than changing the beam-down operator. In this case the beam-down operator is a fixed logic $\mathcal{B}_{ps}$. We can extend the query q as follows:

Definition 6.17 Given a search path $S_1 \rightsquigarrow \ldots \rightsquigarrow S_k$ and a corresponding search path logic $\mathcal{A}_{ps}$ then the *expanded query* S_q is defined as:

$$S_q = \bigcup \{T \mid \vdash_{\mathcal{A}_{ps}} T \,\Box\!\!\rightsquigarrow S_k\}.$$

In words, the expanded query is the union of all T which are about the last node of the search path, in terms of the search path logic. For instance, if the search path logic is $\langle \mathcal{L}, \emptyset, \{\text{Left Singleton Monotonic Union}\}\rangle$ and S_k the last node of the search path then $S_q = S_k \cup \{\phi \mid S_k \cup \{\phi\} \in N_i\}$.

In the following section we will be less strict in proposing one unique aboutness proof system or one single end situation S_q. We postulate that the ordering techniques of Section 6.2 could be used to propose an ordering on aboutness aboutness proof systems obtained from the search path. This ordering is based on the way in which the user follows the links through the hyperindex.

6.3.4 Ordering of aboutness proof systems

In this section we show that given a search path we are able to infer a preference relation over aboutness proof systems in the way it is presented in Section 6.2. Here we propose that given a search path we can infer a set of aboutness proof systems and an ordering function π over this set. As there are several logics involved, we may consider various ordering functions. Let us start with the most obvious one, the beam-down logic as it is created after inspecting the search path. We will define a function SPS_{ps} that given a search path results in a set of aboutness proof systems.

Definition 6.18 Let P be a search path $S_1 \leadsto \ldots \leadsto S_k$. The function SPS_{ps} is defined as follows: $\mathrm{SPS}_{ps}(P) = [\mathrm{SP}_{ps}(\mathcal{A}_1), \ldots, \mathrm{SP}_{ps}(\mathcal{A}_{k-1})]$ with $\mathcal{A}_i$ the search path logic of the search path $S_{k-i} \leadsto S_{k+1-i}$.

Here π can be defined by $\pi(i) = i$. Let us explain the intuition behind this definition. A search path is evaluated under the assumption that the last steps are more important than the first steps. Then the corresponding aboutness proof system of the last step is more important than the corresponding aboutness proof system of the last two steps etc. The resulting preference states that documents retrieved with an aboutness proof system of the last step are preferred over the documents retrieved with an aboutness proof system of the last two steps.

For instance, if the last step was based on thesaurus information, a document that is retrieved using a thesaurus for the final query is preferred over, for instance, more specific document representation.

In the work of Berger [13, 14, 16, 17] another use of evaluating the search path is suggested, the so-called *search support*. Such a navigation aid should make suggestions to the user as to which of the nodes of the hyperindex will most likely lead to the search target. In our framework we propose a ranked list of situations. On top of this list is the situation which covers the information where the user is searching for probably better than the lower ones.

The story is similar to the ordering function of the beam-down derivation postulates. Given a user arrived at node S_k after she travelled by search path P. All the situations which are about S_k, using the first aboutness proof system of the list $\mathrm{SPS}_{ps}(P)$ are presented above the list of new reachable nodes. Then followed by all situations which are about S_k using the second aboutness proof system of the list $\mathrm{SPS}_{ps}(P)$, and so on.

The user is guided as the system delivers her an ordering based on her previous search actions, after every search action the preferences of the next nodes changes.

6.3.5 Defaults in a query-by-navigation process

In Chapter 5 we introduced the user's defaults which, if available, could be of great help in order to increase precision. Let us consider the example given in Chapter 5 were we used the query "programs" and two documents "computer programs" and "television programs". In a hypertext environment the user starts at the node $\{\langle\langle\text{Programs}\rangle\rangle\}$. At this point she can decide to choose the computer- or the television-interpretation. If she wishes to be informed about "computer programs", she probably chooses the node $\{\langle\langle\text{Programs}\rangle\rangle, \langle\langle\text{Computer}\rangle\rangle\}$ and by travelling through the hyperindex never reaches nodes related to television-aspects. We can derive a default that most likely the information need of the user is not related to television.

Definition 6.19 Given a hypertext $\mathrm{I} = \langle N_i, E_i\rangle$ and a corresponding hypertext logic $\mathcal{A}_{ps}$. Furthermore let a search path P be given. If $S_j \rightsquigarrow S_j \cup \{\varphi\}$ is a part of the search path then the beam-down logic $\mathcal{B}_{ps}$ can be extended with the axiom $\psi \bot \varphi$ if and only if for all situations S_i in search path P, $\psi \notin S_i$ and $S_i \cup \{\psi\} \in N_i$.

In our example, the user went from the node $\{\langle\langle\text{Programs}\rangle\rangle\}$ to the node $\{\langle\langle\text{Programs}\rangle\rangle, \langle\langle\text{Computer}\rangle\rangle\}$, although she might as well have walked to the node $\{\langle\langle\text{Programs}\rangle\rangle, \langle\langle\text{Television}\rangle\rangle\}$. If she does not considers a node with the profon $\langle\langle\text{Television}\rangle\rangle$ then we can on the basis of this information adopt the default that information about computers precludes information about televisions. Note that this kind of defaults are purely based on the user's actions, not on statistical information, neither on general defaults.

6.3.6 Conclusion

One of the advantages of a theoretical approach for modelling an integrated information retrieval hypermedia model is that one can study aboutness decisions as they occur in the model. We formalised three different kinds of logics, a hyperindex logic, a beam-down logic, and a search path logic. In our approach the construction of these three logics could be dependent of each other. A search path logic could be used to influence the beam-down operator or for query expansion. The way a user travelled through the hyperindex could be used to propose an ordering of nodes of the hyperbase as well as of the hyperindex. Finally we presented briefly that the search path logic could be used to generate some defaults which can be used for improving the representation of the user's information need.

6.4 Summary and conclusions

In this chapter we have presented the impact of our framework for studying information retrieval by proposing some properties, techniques, and ideas for combining and ordering aboutness proof systems. There are many avenues for further research. To begin with, a further investigation of useful concepts, definitions, and theorems is needed. The definitions of filtering, ordering and so on given in this chapter are just a first start of the theoretical study of IR models. Detailed investigation of more definitions and theorems are needed in order to describe a complete information retrieval theory. Such a complete information retrieval theory must accurately predict the results of possible IR models or combinations of it.

Furthermore, this section presented a formalisation of a two-level hypermedia approach, include a so-called query-by-navigation process. The framework is based on our information retrieval theory in which it is possible to study 'hyperindex modelling'-related questions. The ordering of the documents is influenced by the way the user follows links to the hyperindex. Certain decisions have certain consequences. In our proposed framework, we are able to express the decisions and their consequences for the ordering of the documents. The user is no longer a *'to be forgotten'* object. She can play a pivotal role in the ordering of documents using the query-by-navigation process. Of course there are interesting aspects we did not study closely. Pertinent questions include: 'what is the role of the information links between documents in the document-base' and 'how can we use the profiles of different users with the same *style*'. What we did show was that it is possible to infer what the consequences are of certain decisions. This can be useful for those who want to build new hypermedia systems or extended existing ones or those who want to compare different systems based on the way the links in the systems are defined.

Chapter 7

Conclusions and future work

I may be wrong and you may be right,
and by an effort, we may get nearer to the truth.

K.R. Popper, *'The Open Society and Its Enemies'*.

In this final chapter we start with a discussion of the main results and achievements of this thesis. We also point out some opportunities for future research provided by the axiomatic theory proposed in the preceding chapters. The chapter ends with some sketches for more general practical and theoretical research, as it could be carried out in information retrieval as well as in other areas.

7.1 Overview

In Chapter 1 we sketched the information retrieval paradigm. We showed the different approaches for modelling and studying information retrieval. Next, in Chapter 2 the reasons and consequences of proposing an information theory are given. Situation Theory as a meta-language is introduced in Chapter 3, which furthermore contains the formalisation of aboutness proof systems. In Chapter 4 we investigated several common information retrieval models as proposed in the academic world. By mapping these models into our framework, we were able in Chapter 5 to analyse and compare them theoretically. In Chapter 6 three elaborate examples are given that show how our framework can be used. The first example shows how IR systems can be combined based on qualitative grounds. In the second example it is shown how the output of an IR system can be ordered based on a preference based on axiomatic definitions of aboutness. In the third example, a two-level hypermedia application is proposed in a theoretically well-founded way.

7.2 Achievements

In this thesis we have presented a framework that allows us to model various kinds of IR models. We showed that one can prove that the theoretical models are indeed correct representations of the IR models, by proving appropriate soundness and completeness theorems. The fact that many important IR models can be characterised by means of (sound and complete) aboutness proof systems allowed us to compare the models theoretically instead of experimentally. We showed equivalence and embedding relations between IR models. These results allow us to sidestep the controversies surrounding the experimental comparison of IR models. We also showed that some models are monotonic whereas others are not. In case of monotonicity of an IR model, we were able to make qualitative statements about the recall values obtained by the model. Since some information retrieval researchers argue that IR models should display a non-monotonic character, we presented some rules that transfer a monotonic aboutness proof system into a non-monotonic one. Furthermore, we gave an extensive list of postulates useful for information retrieval.

7.3 Future research

7.3.1 Information retrieval

The first extension would be the mapping of several other models. We believe that every IR model deriving aboutness in a more or less logical sense can be translated to an aboutness proof system in our framework. Another extension could be to create a method which makes it possible to build a cognitively acceptable aboutness relation from a given set of postulates. Although the expressive power of the framework has been demonstrated, we did not pay attention to aspects related to computational complexity.

Another problem to pursue concerns extensions to the formalisation of anti-aboutness. Suitable extensions could be used for information filtering. At first, an aboutness proof system derives all documents which are about the query. With a detailed formalisation of anti-aboutness, it is possible to transfer anti-aboutness information on the retrieved set of documents. This last step could be viewed as an information filtering process.

7.3.2 Situation Theory

In the axiomatic theory the representatives of the information content of both the document and of the information need are situations. The presented IR models are using a rather simplistic representation of information in which features such as context-representation, nested information, and backgrounds are lacking (for an overview of some essential features an information retrieval representation should have see [85]). Therefore the full expressive power of Situation Theory for information retrieval is neither used nor shown in this thesis. However, in her thesis [84] Lalmas showed that Situation Theory covers all these features.

A more complex representation of documents and queries in combination with the axiomatic theory could lead us to a better understanding of the nature of information in information retrieval (see [69] where this conviction is explained more comprehensively). The nature of information is manifold and can be studied from different perspectives. For example, in future research, we plan to look at the nature of user modelling and its influence on the notion of aboutness. A correct representation of the user's (mental) intention will probably generate better retrieval.

We also want to study some aspects of information modelling that involve logical problems, like inconsistency, paradoxes, and tautologies. Situation Theory allows us to tackle these (see for example [7] in which Barwise & Etchemendy are modelling paradoxes using Situation Theory). The use of a situation-theoretical representation of information in combination with our axiomatic theory of aboutness is an interesting avenue for further investigation.

Another issue is an extension of the set of postulates in such a way that the notion of aboutness becomes context-dependent. The representation of contexts, which recently has reached higher prominence in information retrieval, will be investigated using Situation Theory. For example, with network information retrieval (or any distributed database) it is necessary to represent the fact that retrieval is with respect to a specific site (a context). Moreover, two sites (contexts) may be involved in an aboutness derivation and the information retrieved from them must be aggregated. Using Situation Theory it is possible to capture this notion of context using situations and background conditions.

Finally, we want to look at the possibility of using Channel Theory. In [127], an IR model is developed based on Channel Theory, a novel approach based on Situation Theory [5, 6] in which the nature of the information flow can be defined by constraints between pairs of situations. The information flow is said to be carried by a channel. In the Channel-theoretic approach, the document and the query are represented by situations. Determining the relevance of a document is to find the channel, together with its nature, that led from the situation modelling the document to the situation containing the information being sought. The channel could be build as the (sequential and/or parallel) combination of more primitive channels. The synthesis of this approach with ours is considered by Lalmas [83] as follows:

> *We believe that the use of channels presents the most potential for IR modelling. For example, a different use of channels is one where a channel models*

> *a retrieval method. Indeed, one can define several types of flows, one for each type of information retrieval methods (Boolean, probabilistic, vector space or logical). A method can be used separately (i.e., one channel is involved) or can be combined with one or more other methods (i.e., parallel channels are involved). The document that is retrieved by many methods can be considered to be highly relevant to the information need. Obviously, it is necessary to define what a Boolean or a vector space flow is. The advantage of this approach is that, as well as being able to model different IR methods, the model can be used to compare them formally. The properties of the corresponding flows might lead to interesting results.*

In this light we could see our work as a study of the flow of information, or aboutness, which can be modelled as a channel. Then, an R-system could be represented as a typical channel, as well as a C-system, or a vector-space aboutness proof system, and so on. Further research is necessary to analyse the approach in more detail.

7.3.3 Databases

In Chapter 1, we summarised the difference between data retrieval and information retrieval. The theory presented in this thesis could possibly be a step towards an integration of database systems and IR systems. In database systems, aboutness is defined in terms of an R-system in which a fact is about another fact if both facts are the same. To extend this definition in order to allow plausible inference, some of the postulates presented in Chapter 3 can be chosen.

7.3.4 Artificial Intelligence

As mentioned in Chapter 1, research in information retrieval and research in artificial intelligence are more and more converging. From the point of view of information retrieval, people are interested in all types of non-monotonic reasoning. Furthermore, knowledge representation languages developed in AI (for example Conceptual Graphs [34] and Terminological Logic [101]) are inspected for their possible use in information retrieval.

Meanwhile, theoretical research in AI is searching for a so-called ‘killer application’ that could show the usefulness of a specific theory/formalism/language/approach. Information retrieval is often considered to be a possible area of practical application of such theoretical approaches, since the fundamentals of information retrieval and AI, such as information, inference, uncertainty, and so on, are very close related.

The theory presented in this thesis was based on the meta-theory of Kraus, Lehman & Magidor [81]. The main interest of Kraus, Lehman & Magidor was to study non-monotonic relations, whereas our interest is the study of aboutness relations. It would be interesting to compare these two meta-theories in order to see where they differ and whether these differences are intuitively acceptable and explainable. Or stated differently, are there some typical aboutness properties which are not reasoning properties and vice versa and more importantly, can we formulate an explanation for these differences in terms of our concept of aboutness. This analysis could lead to an improvement of our theory of aboutness.

Finally we hope that information retrieval researchers can benefit from the theory in such a way that the theory can describe relevance decisions in a cognitively acceptable way and that it can be used to predict the change in behaviour due to a modification of the logical essence of an IR model.

Bibliography

[1] IJ.J. Aalbersberg. A document retrieval model based on term frequency ranks. In W. Bruce Croft and C.J. van Rijsbergen, editors, *Proceedings of the Seventeenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 163–172, Dublin, July 1994. Springer-Verlag, Berlin.

[2] P. Aczel, D. Israel, Y. Katagiri, and S. Peters, editors. *Situation Theory and its Applications, Volume 3*, CSLI Lecture Notes, Number 37. CSLI, Stanford, 1993.

[3] M. Agosti, R. Colotti, and G. Gradenigo. A two-level hypertext retrieval model for legal data. In A. Bookstein, Y. Chiaramella, G. Salton, and V.V. Raghavan, editors, *Proceedings of the Fourteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 316–325, Chicago, June 1991. ACM Press, New York.

[4] J. Barwise. *The Situation in Logic.* CLSI Lecture Notes, Number 17. CSLI, Stanford, 1989.

[5] J. Barwise. Information links in domain theory. In S. Brookes, M. Main, A. Melton, M. Mislove, and D. Schmidt, editors, *Proceedings of the Mathematical Foundation of Programming Semantics Conference*, Lecture Notes in Computer Science, Number 598, pages 168–192. Springer-Verlag, Berlin, 1991.

[6] J. Barwise. Constraints, channels and the flow of information. In P. Aczel, D. Israel, Y. Katagiri, and S. Peters, editors, *Situation Theory and its Applications, Volume 3*, CSLI Lecture Notes, Number 37, pages 3–27. CSLI, Stanford, 1993.

[7] J. Barwise and J. Etchemendy. *The Liar, An Essay on Truth and Circularity.* Oxford University Press, Oxford, 1987.

[8] J. Barwise and J. Etchemendy. Information, infons, and inference. In R. Cooper, K. Mukai, and J. Perry, editors, *Situation Theory and its Applications, Volume 1*, CSLI Lecture Notes, Number 22, pages 33–78. CSLI, Stanford, 1990.

[9] J. Barwise, J.M. Gawron, G. Plotkin, and S. Tutiya, editors. *Situation Theory and its Applications, Volume 2*, CSLI Lecture Notes, Number 26. CSLI, Stanford, 1991.

[10] J. Barwise and J. Perry. Situations and attitudes. *Journal of Philosophy*, 78(11):668–691, 1981.

[11] J. Barwise and J. Perry. *Situations and Attitudes.* Bradford Book, MIT Press, Cambridge, Massachusetts, 1983.

[12] H. van den Berg. *Knowledge Graphs and Logic, One of Two Kinds.* PhD thesis, Department of Applied Mathematics, Universiteit Twente, The Netherlands, September 1993.

[13] F.C. Berger. Query construction via navigation. Technical Report CSI-R9514, Computing Science Institute, University of Nijmegen, Nijmegen, The Netherlands, November 1995.

[14] F.C. Berger, A.H.M. ter Hofstede, and T.P. van der Weide. Supporting query by navigation. In R. Leon, editor, *Information retrieval: New systems and current research -Proceedings of the 16th Research Colloquium of the British Computer Society Information Retrieval Specialist Group-*, pages 26 – 46, Drymen, Scotland, March 1994. Taylor Graham, London.

[15] F.C. Berger and T.W.C. Huibers. A framework based on situation theory for searching in a thesaurus. *The New Review of Document and Text Management*, 1:253–276, 1995.

[16] F.C. Berger and P. van Bommel. Personalized Search Support for Networked Document Retrieval Using Link Inference. In R.R. Wagner and H. Thoma, editors, *Proceedings of the 7th International Conference DEXA'96 on Data Base and Expert System Applications Conference*, volume 1134, pages 802–811, Zurich, Switzerland, September 1996. Springer-Verlag, Berlin.

[17] F.C. Berger and T.P. van der Weide. A feedback mechanism for query by navigation. In R. Sacks-Davis and J. Zobel, editors, *Proceedings of the Sixth Australasian Database Conference, ADC'95*, volume 17(2) of *Australian Computer Science Communications*, pages 56–65, Adelaide, January 1995.

[18] C. Berrut. *Une méthode d'indexation fondée sur l'analyse sémantique de documents spécialisés. Le prototype RIME et son application à un corpus médical.* PhD thesis, Laboratoire de Génie Informatique, Université Joseph Fourier, Grenoble, France, December 1988.

[19] D.C. Blair. *Language and Representation in Information Retrieval.* Elsevier Science Publishers, Amsterdam, 1990.

[20] L. de Brabandere. *Les Infoducs.* Editions Duculot, Gembloux, 1985.

[21] F.P. Brooks and K.E. Iverson. *Automatic Data Processing.* John Wiley & Sons, Inc., New York, 1963.

[22] W. Bruce Croft and C.J. van Rijsbergen, editors. *Proceedings of the Seventeenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Dublin, July 1994. Springer-Verlag, Berlin.

[23] P.D. Bruza. *Stratified Information Disclosure, a Synthesis between Hypermedia and Information Retrieval.* PhD thesis, University of Nijmegen, The Netherlands, March 1993.

[24] P.D. Bruza and L.C. van der Gaag. Efficient context-sensitive plausible inference for information disclosure. In R. Korfhage, E. Rasmussen, and P. Willett, editors, *Proceedings of the Sixteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 12–21, Pittsburgh, June 1993. ACM Press, New York.

[25] P.D. Bruza and T.W.C. Huibers. Detecting the erosion of hierarchic information structures. In M. Murata and H. Gallaire, editors, *Proceedings of Principles of Document Processing '94*, Darmstadt, Germany, April 1994.

[26] P.D. Bruza and T.W.C. Huibers. Investigating aboutness axioms using information fields. In W. Bruce Croft and C.J. van Rijsbergen, editors, *Proceedings of the Seventeenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 112–121, Dublin, July 1994. Springer-Verlag, Berlin.

[27] P.D. Bruza and T.W.C. Huibers. How nonmonotonic is aboutness? Technical Report UU-CS-1995-09, Department of Computer Science, Utrecht University, The Netherlands, March 1995.

[28] P.D. Bruza and T.W.C. Huibers. A study of aboutness in information retrieval. *Artificial Intelligence Review*, 10(5-6):381–407, 1996.

[29] P.D. Bruza and T.P. van der Weide. A two level hypermedia - an improved architecture for hypertext. In A.M. Tjoa and R. Wagner, editors, *Proceedings of the Data Base and Expert System Applications Conference (DEXA 90)*, pages 76–83. Springer-Verlag, Berlin, 1990.

[30] M. Buckland and F. Gey. The relationship between recall and precision. *Journal of the American Society for Information Science*, 45(1):12–19, January 1994.

[31] J.P. Callan, W. Bruce Croft, and S.M. Harding. The INQUERY retrieval system. In *Proceedings of the Third International Conference on Database and Expert Systems Applications (DEXA 92)*, pages 78–83, Valencia, Spain, 1992. Springer-Verlag, Berlin.

[32] J.P. Callan, Z. Lu, and W. Bruce Croft. Searching distributed collections with inference networks. In E.A. Fox, P. Ingwersen, and R. Fidel, editors, *Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 21–28, Seattle, July 1995. ACM, ACM Press, New York.

[33] A. Cawsey, G. Galliers, S. Reece, and K. Sparck Jones. Automating the librarian: Belief revision as a base for system action and communication with the user. *The Computer Journal*, 35(3):221–232, 1992.

[34] J.-P. Chevallet. *Un modèle logique de recherche d'informations appliqué au formalisme des graphes conceptuels. Le prototype ELEN et son expérimentation sur un corpus de composants logiciels*. PhD thesis, Laboratoire de Génie Informatique, Université Joseph Fourier, Grenoble I, France, May 1992.

[35] Y. Chiaramella and J.-P. Chevallet. About retrieval models and logic. *The Computer Journal*, 35(3):233–241, 1992.

[36] C.W. Cleverdon. Comparative efficiency of indexing systems. 1960–1962. 2 Volumes.

[37] C.W. Cleverdon. The significance of the Cranfield tests on index languages. In A. Bookstein, Y. Chiaramella, G. Salton, and V.V. Raghavan, editors, *Proceedings of the Fourteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 3–12, Chicago, October 1991. ACM, ACM Press, New York.

[38] F.R. Compagnoni and K. Erlich. Information retrieval using a hypertext-based help system. In N.J. Belkin and C.J. van Rijsbergen, editors, *Proceedings of the Twelfth International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 212–220, Cambridge, Massachusetts, 1989. ACM Press, New York.

[39] R. Cooper, K. Mukai, and J. Perry, editors. *Situation Theory and its Applications, Volume 1*, CSLI Lecture Notes, Number 22. CSLI, Stanford, 1990.

[40] W.S. Cooper. Expected search length: A single measure of retrieval effectiveness based on the weak ordering action of retrieval systems. *American Documentation*, pages 30–41, 1968.

[41] W.S. Cooper. A definition of relevance for information retrieval. *Information Storage and Retrieval*, 7:19–37, 1971.

[42] W.S. Cooper. Some inconsistencies and misnomers in probabilistic information retrieval. In A. Bookstein, Y. Chiaramella, G. Salton, and V.V. Raghavan, editors, *Proceedings of the Fourteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 57–61, Chicago, October 1991.

[43] W.S. Cooper. The formalism of probability theory in IR: A foundation for an encumbrance? In W. Bruce Croft and C.J. van Rijsbergen, editors, *Proceedings of the Seventeenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 242–247, Dublin, July 1994. Springer-Verlag, Berlin.

[44] W.S. Cooper. Some inconsistencies and misidentified modeling assumptions in probabilistic information retrieval. *ACM Transactions on Information Systems*, 13(1):100–111, January 1995.

[45] F. Crestani. Probability kinematics in information retrieval. In E.A. Fox, P. Ingwersen, and R. Fidel, editors, *Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 291–299, Seattle, July 1995. ACM Press, New York.

[46] F. Crestani and C.J. van Rijsbergen. Information retrieval by imaging. In R. Leon, editor, *Information retrieval: New systems and current research -Proceedings of the 16th Research Colloquium of the British Computer Society Information Retrieval Specialist Group-*, pages 47 –67, Drymen, Scotland, March 1994. Taylor Graham, London.

[47] J. December. Challenges for web information providers. *Computer-Mediated Communication Magazine*, 1(6):8 – 14, October 1994.

[48] K. Devlin. Infons and types in an information-based logic. In R. Cooper, K. Mukai, and J. Perry, editors, *Situation Theory and its Applications, Volume 1*, CSLI Lecture Notes, Number 22, pages 79–95. CSLI, Stanford, 1990.

[49] K. Devlin. *Logic and Information.* Cambridge University Press, Cambridge, England, 1991.

[50] F.I. Dretske. *Knowledge and the Flow of Information.* Basic Blackwell Publisher, 1981.

[51] J. Farradane. Relational indexing, part I. *Journal of Information Science*, 1(5):267–276, 1980.

[52] J. Farradane. Relational indexing, part II. *Journal of Information Science*, 1(6):313–324, 1980.

[53] E.A. Fox, P. Ingwersen, and R. Fidel, editors. *Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Seattle, July 1995. ACM Press, New York.

[54] N. Fuhr. Probabilistic models in information retrieval. *The Computer Journal*, 35(3):243–255, 1992.

[55] N. Fuhr. Probabilistic datalog – a logic for powerful retrieval methods. In E.A. Fox, P. Ingwersen, and R. Fidel, editors, *Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 282–290, Seattle, July 1995. ACM Press, New York.

[56] D. Gabbay. Labelled deductive systems and situation theory. In P. Aczel, D. Israel, Y. Katagiri, and S. Peters, editors, *Situation Theory and its Applications, Volume 3*, CSLI Lecture Notes, Number 37, pages 89–118. CSLI, Stanford, 1993.

[57] L.T.F. Gamut. *Introduction to Logic*, volume 1 of *Logic, Language, and Meaning.* The University of Chicago Press, Chicago, 1991.

[58] D. Haines and W. Bruce Croft. Relevance feedback and inference networks. In R. Korfhage, E. Rasmussen, and P. Willett, editors, *Proceedings of the Sixteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2–11, Pittsburgh, June 1993. ACM Press, New York.

[59] L. Hardman, D.C.A. Bulterman, and G. van Rossum. The amsterdam hypermedia model: Adding time and context to the dexter model. *Communications of the ACM*, 37(2):50–62, 1994.

[60] D.K. Harman, editor. *The first Text REtrieval Conference (TREC-1)*, NIST Special Publication 500–207, Gaithersburg MD, 1993.

[61] D.K. Harman. Overview of the third Text REtrieval Conference (TREC-3). In D.K. Harman, editor, *The third Text REtrieval Conference (TREC-3)*, pages 1–20, Gaithersburg, Maryland, 1994. ACM Press.

[62] D.K. Harman, editor. *The second Text REtrieval Conference (TREC-2)*, NIST Special Publication 500–215, Gaithersburg MD, 1994.

[63] D.K. Harman, editor. *The third Text REtrieval Conference (TREC-3)*, NIST Special Publication 500–255, Gaithersburg MD, 1994.

[64] S. Hawking. *A Brief History of Time, From the Big Bang to Black Holes.* Bantam Books, London, 1988.

[65] G.E. Hughes and M.J. Cresswell. *A Companion to Modal Logic.* Methuen & Co. Ltd., London, 1984.

[66] T.W.C. Huibers. Towards an axiomatic aboutness theory for information retrieval. In F. Crestani and M. Lalmas, editors, *Proceedings of the 2nd Workshop on Information Retrieval, Uncertainty and Logic (WIRUL '96)*, Glasgow, July 1996. Electronic version.

[67] T.W.C. Huibers and P.D. Bruza. Situations, a general framework for studying information retrieval. In R. Leon, editor, *Information retrieval: New systems and current research -Proceedings of the 16th Research Colloquium of the British Computer Society Information Retrieval Specialist Group-*, pages 3–24, Drymen, Scotland, March 1994. Taylor Graham, London.

[68] T.W.C. Huibers and N. Denos. A qualitative ranking method for logical information retrieval models. In M. Lalmas, editor, *Proceedings of the Workshop on the treatment of Uncertainty in Logic-based Models of Information Retrieval Systems*, Glasgow, September 1995. Electronic version (also appeared as Technical Report RAP95-005, Groupe MRIM of the Laboratoire de Génie Informatique, Grenoble, France).

[69] T.W.C. Huibers, M. Lalmas, and C.J. van Rijsbergen. Information retrieval and situation theory. *SIGIR Forum*, 30(1):11–25, 1996.

[70] T.W.C. Huibers and B. van Linder. Formalising intelligent information retrieval agents. In F. Johnson, editor, *Proceedings of the 18th BCS IRSG Annual Colloquium on Information Retrieval Research*, pages 125–143, Manchester, March 1996.

[71] T.W.C. Huibers, B. van Linder, and P.D. Bruza. Een theorie voor het bestuderen van information retrieval modellen. In L.G.M. Noordman and W.A.M. de Vroom, editors, *Informatiewetenschap 1994: Wetenschappelijke Bijdragen aan de Derde StinfoN Conferentie*, pages 85 – 102, Tilburg, the Netherlands, December 1994. Stichting StinfoN. (In Dutch).

[72] T.W.C. Huibers, B. van Linder, and J.-J. Ch. Meyer. An agent-oriented approach to information retrieval. To appear in proceedings of NAIC'96, Utrecht, The Netherlands, 1996.

[73] T.W.C. Huibers, I. Ounis, and J.-P. Chevallet. Axiomatization of a conceptual graph formalism for information retrieval in a situated framework. Technical Report RAP95-004, Groupe MRIM of the Laboratoire de Génie Informatique in Grenoble, France, July 1995.

[74] T.W.C. Huibers, I. Ounis, and J.-P. Chevallet. Conceptual graph aboutness. In P.W. Eklund, G. Ellis, and G. Mann, editors, *Conceptual Structures: Knowledge Representation as Interlingua, 4th International Conference on Conceptual Structures (ICCS'96)*, volume 1115 of *Lecture Notes in Artificial Intelligence*, pages 130 – 144, Sydney, August 1996. Springer-Verlag, Berlin.

[75] D. Hull. Using statistical testing in the evaluation of retrieval experiments. In R. Korfhage, E. Rasmussen, and P. Willett, editors, *Proceedings of the Sixteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 329–338, Pittsburgh, 1993. ACM Press, New York.

[76] A. Hunter. Using default logic in information retrieval. In C. Froidevaux and J. Kohlas, editors, *Symbolic and Quantitative Approaches to Uncertainty*, number 946 in Lecture Notes in Computer Science, pages 235–242. Springer-Verlag, Berlin, 1995.

[77] J.J. IJdens. Using index expression belief networks for information disclosure. Master's thesis, Department of Computer Science, Utrecht University, The Netherlands, March 1994.

[78] A. Kheirbek. *Modèle d'intégration d'un Système de Recherche d'Informations et d'un Système Hypermédia basé sur le formalsime des Graphes Conceptuels.* PhD thesis, Laboratoire de Génie Informatique,Université Joseph Fourier - Grenoble I, France, May 1995.

[79] A. Kheirbek and Y. Chiaramella. Integrating hypermedia and information retrieval with conceptual graphs formalism. In *Proceedings of the Hypertext -Information Retrieval- Multimedia Conference HIM'95*, pages 47 – 60, Konstanz, April 1995. Universitatsverlag Konstanz.

[80] R. Korfhage, E. Rasmussen, and P. Willett, editors. *Proceedings of the Sixteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Pittsburgh, June 1993. ACM Press, New York.

[81] S. Kraus, D. Lehmann, and M. Magidor. Nonmonotonic reasoning, preferential models and cumulative logic. *Artificial Intelligence*, 44:167–207, 1990.

[82] M. Lalmas. From a qualitative towards a quantitative representation of uncertainty on a situation theory based model of an information retrieval system. In M. Lalmas, editor, *Proceedings of the Workshop on the treatment of Uncertainty in Logic-based Models of Information Retrieval Systems*, Glasgow, September 1995. Electronic version.

[83] M. Lalmas. The flow of information in information retrieval: its modelling. In F. Crestani and M. Lalmas, editors, *Proceedings of the 2nd Workshop on Information Retrieval, Uncertainty and Logic (WIRUL '96)*, Glasgow, July 1996. Electronic version.

[84] 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, Scotland, April 1996.

[85] M. Lalmas. The use of logic in information retrieval modelling. Departmental Research Report IR-96-1, Department of Computing Science, University of Glasgow, Scotland, January 1996.

[86] M. Lalmas and C.J. van Rijsbergen. A logical model of information retrieval based on situation theory. In *Proceedings of the BCS 14th Information Retrieval Colloquium*, pages 1–13, Lancaster, April 1992. British Computer Society, Springer-Verlag, London.

[87] 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 V.S. Alagar, S. Berger, and F. Dong, editors, *Incompleteness and Uncertainty in Information Systems*, pages 62–67, 1993.

[88] F.W. Lancaster. *Toward Paperless Information Systems.* Academic Press, New York, 1978.

[89] F. Landman. *Towards a Theory of Information. The Status of Partial Objects in Semantics.* Foris, Dordrecht, 1986.

[90] B. van Linder. *Modal Logics for Rational Agents.* PhD thesis, Department of Computer Science, Utrecht University, The Netherlands, June 1996.

[91] B. Logan, S. Reece, and K. Sparck Jones. Modelling information retrieval agents with belief revision. In W. Bruce Croft and C.J. van Rijsbergen, editors, *Proceedings of the Seventeenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 142–151, Dublin, July 1994. Springer-Verlag, Berlin.

[92] D. Lucarella. A model for hypertext-based information retrieval. In *Proceedings of the European Conference on Hypertext - ECHT 90*, pages 81–94. Cambridge University Press, Cambridge, England, 1990.

[93] D. Lucarella and Z. Zanzi. Information retrieval from hypertext: An approach using plausible inference. *Information Processing & Management*, 29(3):299–312, 1993.

[94] W. Łukaszewicz. *Non-Monotonic Reasoning*. Ellis Horwood Series in Artificial Intelligence. Ellis Horwood, New York, 1990.

[95] M.E. Maron. On indexing, retrieval and the meaning of about. *Journal of the American Society for Information Science*, pages 38–43, January 1977.

[96] J. Martin. *Design of real-time computer systems*. Series in Automatic Computation. Prentice Hall Inc., Englewood Cliffs, New Jersey, 1967.

[97] L.J. Matthijssen. An intelligent interface for legal databases. In *International Conference on Artificial Intelligence and Law 1995 Proceedings*, pages 71–80, College Park, Maryland, May 1995. ACM, New York.

[98] L.J. Matthijssen. Information retrieval for legal argumentation tasks, an architecture for legal information retrieval using taskmodels. In *Proceedings of the Fifth National/First European Conference on Law, Computers and Artificial Intelligence*, pages 117–130, Exeter University Centre for Interdisciplinary Legal Studies, April 1996.

[99] M. Mechkour. *EMIR2. Un Modèle étendu de Représentation et de Correspondance d'images pour la Recherche d'informations. Application à un Corpus d'images Historiques.* PhD thesis, Laboratoire de Génie Informatique,Université Joseph Fourier, Grenoble I, France, November 1995.

[100] C. Meghini. An image retrieval model based on classical logic. In E.A. Fox, P. Ingwersen, and R. Fidel, editors, *Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 300–308, Seattle, July 1995. ACM Press, New York.

[101] C. Meghini, F. Sebastiani, U. Straccia, and C. Thanos. A model of information retrieval based on a terminological logic. In R. Korfhage, E. Rasmussen, and P. Willett, editors, *Proceedings of the Sixteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 298–307, Pittsburgh, June 1993. ACM Press, New York.

[102] M.L. Mugnier. On specialization and projection for conceptual graphs. Technical report no. 93-003, Laboratoire d'Informatique, de Robotique et de Microélectronique de Montpellier, France, January 1993.

[103] A. Muller. Abductive retrieval of structured documents. In M. Murata and H. Gallaire, editors, *Proceedings of Principles of Document Processing '94*, Darmstadt, Germany, April 1994.

[104] A. Muller and S. Kutschekmanesch. Using abductive inference and dynamic indexing to retrieve multimedia sgml documents. In I. Ruthven, editor, *MIRO '95, Proceedings of the Final Workshop on Multimedia Information Retrieval*, Glasgow, September 1995. Electronic Workshops in Computing. Springer-Verlag, Berlin. Electronic version.

[105] N. Negroponte. *Being Digital.* Hodder and Stoughton, London, 1995.

[106] J.-Y. Nie. An information retrieval model based on modal logic. *Information Processing & Management*, 25(5):477–491, 1989.

[107] J.-Y. Nie. *Un modèle logique général pour les systèmes de recherche d'informations - Application au prototype RIME.* PhD thesis, Laboratoire de Génie Informatique,Université Joseph Fourier, Grenoble I, France, 1990.

[108] J.-Y. Nie. Towards a probabilistic modal logic for semantic-based information retrieval. In N. Belkin, P. Ingwersen, and A.M. Pejtersen, editors, *Proceedings of the Fifteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 140–151, Copenhagen, June 1992. ACM Press, New York.

[109] J.-Y. Nie and Y. Chiaramella. A retrieval model based on an extended modal logic and its applications to the RIME experimental approach. In J. Vidick, editor, *Proceedings of the 13th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 25–43, Bruxelles, 1990. ACM Press, New York.

[110] J. Nielsen. The art of navigating through Hypertext. *Communications of the ACM*, 33(3):296–310, March 1990.

[111] F. van Oostrom. *De Waarde van het Boek.* Amsterdam University Press, Amsterdam, 1994. (In Dutch).

[112] I. Ounis. Logique terminologique pour la correspondance entre graphes de concepts dans le cadre d'un système de recherche d'informations. Master's thesis, Laboratoire de Génie Informatique,Université Joseph Fourier, Grenoble, France, June 1994.

[113] I. Ounis. Une dénotation pour les graphes conceptuels: comparaison avec les logiques terminologiques en recherche d'informations. In *XIII Congrès INFORSID*, Grenoble, May-June 1995.

[114] I. Ounis and J.-P. Chevallet. Using Conceptual Graphs in a Multifaceted Logical Model for Information Retrieval. In R.R. Wagner and H. Thoma, editors, *Proceedings of the 7th International Conference DEXA'96 on Data Base and Expert System Applications Conference*, pages 812–823, Zurich, Switzerland, September 1996. Springer-Verlag, Berlin.

[115] J. Pearl. *Probabilistic Reasoning in Intelligent Systems: Network of Plausible Inference.* Morgan Kaufmann, San Mateo, California, revised second printing edition, 1991.

[116] C.S. Peirce. *Collected Papers of C.S. Peirce.* Harvard University Press, Cambridge, Massachusetts, 1931–1958. 8 vols.

[117] K.R. Popper. *Objective Knowledge, An Evolutionary Appproach.* Clarendon Press, Oxford, 7th edition, 1972.

[118] K.R. Popper. *The Logic of Scientific Discovery.* Routledge, London, second English edition, 1992. Translation of "Logik der Forschung" published in Vienna 1934.

[119] K.R. Popper. *The Myth of the Framework, in Defence of Science and Rationality.* Routledge, London, 1994.

[120] R. Reiter. On closed-world data bases. In H. Gallaire and J. Minker, editors, *Logic and Data Bases*, pages 55–76. Plenum Press, New York, 1978.

[121] R. Reiter. A logic for default reasoning. *Artificial Intelligence*, 13:81–132, 1980.

[122] C.J. van Rijsbergen. *Information Retrieval.* Butterworth & Co (Publishers) Ltd, London, second edition, 1979.

[123] C.J. van Rijsbergen. A new theoretical framework for information retrieval. In F. Rabiti, editor, *Proceedings of the 9th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 194–200, Pisa, Italia, September 1986. ACM, ACM Press, New York.

[124] C.J. van Rijsbergen. A non-classical logic for information retrieval. *The Computer Journal*, 29(6):481–485, 1986.

[125] C.J. van Rijsbergen. Probabilistic retrieval revisited. *The Computer Journal*, 35(3):291–298, 1992.

[126] C.J. van Rijsbergen. Two essays in information retrieval. Departmental Research Report IR-93-3, Department of Computing Science, University of Glasgow, Scotland, November 1993.

[127] 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.

[128] S.E. Robertson. The probability ranking principle in IR. *Journal of Documentation*, 33(4):294–304, 1977.

[129] D. Rogers. *The Bodleian Library and its Treasures 1320–1700.* Aidian Ellis, Oxon, 1991.

[130] B. Russell. *History of Western Philosophy.* George Allen & Unwin Ltd, second edition, 1961. Reprinted version available from Routledge, London.

[131] G. Salton. *The SMART Retrieval System.* Prentice Hall Inc., Englewood Cliffs, New Jersey, 1971.

[132] G. Salton. The state of retrieval system evaluation. *Information Processing & Management*, 28(4):441–449, 1992.

[133] S.M. Schulz. Modal situation theory. In P. Aczel, D. Israel, Y. Katagiri, and S. Peters, editors, *Situation Theory and its Applications, Volume 3*, CSLI Lecture Notes, Number 37, pages 163–188. CSLI, Stanford, 1993.

[134] F. Sebastiani. A probabilistic terminological logic for modelling information retrieval. In W. Bruce Croft and C.J. van Rijsbergen, editors, *Proceedings of the Seventeenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 122–130, Dublin, July 1994. ACM, Springer-Verlag, Berlin.

[135] J. Seligman. Perspectives in situation theory. In R. Cooper, K. Mukai, and J. Perry, editors, *Situation Theory and its Applications, Volume 1*, CSLI Lecture Notes, Number 22, pages 147–192. CSLI, Stanford, 1990.

[136] A.F. Smeaton. *Using parsing of natural language as part of document retrieval.* PhD thesis, University College Dublin, Dublin, Ireland, 1988.

[137] J.F. Sowa. *Conceptual Structures: Information Processing in Mind and Machine.* Addison-Wesley Publishing Company, Reading, MA, 1984.

[138] J.F. Sowa. Relating diagrams to logic. In *Proceedings of the First International Conference in Conceptual Structures, ICCS'93*, volume 699 of *Lecture Notes in Artificial Intelligence*, pages 1–35, Quebec city, August 1993. Springer-Verlag, Berlin.

[139] K. Sparck Jones. Reflections on TREC. *Information Processing & Management*, 31(3):291–314, 1995.

[140] K. Sparck Jones and C.J. van Rijsbergen. Report on the need for and provision of an "ideal" information retrieval test collection. British Library Research and Development Report 5266, Computer Laboratory, University of Cambridge, 1975.

[141] P. Stotts and R. Furuta. Petri-net-based hypertext: Document structure with browsing semantics. *ACM Transactions on Information Systems*, 7(1):3–29, 1989.

[142] J. Tague, A. Salminen, and C. McClellan. Complete formal model for information retrieval systems. In A. Bookstein, Y. Chiaramella, G. Salton, and V.V. Raghavan, editors, *Proceedings of the Fourteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 14–20, Chicago, 1991. ACM Press, New York.

[143] H.R. Turtle and W. Bruce Croft. A comparison of text models. *The Computer Journal*, 35(3):279–290, 1992.

[144] B. Wondergem, W. van der Hoek, T.W.C. Huibers, and C. Witteveen. Preferential semantics for query by navigation. Technical Report CSI-R9616, Computing Science Institute, University of Nijmegen, Nijmegen, The Netherlands, September 1996.

[145] S.K.M. Wong and Y.Y. Yao. On modeling information retrieval with probabilistic inference. *ACM Transactions on Information Systems*, 13(1):36–68, January 1995.

[146] Y.Y. Yao. Measuring retrieval effectiveness based on user preference of documents. *Journal of the American Society for Information Science*, 46(2):133–145, 1995.

[147] E.N. Zalta. Twenty-five basic theorems in situation and world theory. *Journal of Philosophical Logic*, 22(4):385–428, 1993.

Glossary

Axioms

Reflexivity (Re) $S \,\square\!\!\rightsquigarrow S$

Singleton Reflexivity (SR) $\{\varphi\} \,\square\!\!\rightsquigarrow \{\varphi\}$

Rules

Aboutness Inheritance (AI)

$$\frac{S \,\square\!\!\rightsquigarrow_1 T}{S \,\square\!\!\rightsquigarrow_2 T}$$

Careful User (CU)

$$\frac{S \boxtimes\!\!\not\rightsquigarrow_{r_1} T \ldots S \,\square\!\!\rightsquigarrow_{r_i} T \ldots S \boxtimes\!\!\not\rightsquigarrow_{r_k} T}{S \,\square\!\!\rightsquigarrow_{u_3} T}$$

Cautious Monotonicity (CM)

$$\frac{S \,\square\!\!\rightsquigarrow T \qquad S \,\square\!\!\rightsquigarrow U}{S \cup U \,\square\!\!\rightsquigarrow T}$$

Cautious Negation Rationale (CNR)

$$\frac{S \boxtimes\!\!\rightsquigarrow T \quad S \,\square\!\!\not\rightsquigarrow U}{S \boxtimes\!\!\rightsquigarrow T \cup U}$$

Closed World Assumption (CWA)

$$\frac{S \boxtimes\!\!\rightsquigarrow_1 \{\varphi\} \quad \varphi \perp_2 \psi}{S \,\square\!\!\rightsquigarrow_2 \{\psi\}}$$

Composition (Cp)

$$\frac{S \,\square\!\!\rightsquigarrow T}{S \,\square\!\!\rightsquigarrow T \cap U}$$

Composition Preclusion (CP)

$$\frac{S \cup \{\varphi\} \,\square\!\!\rightsquigarrow \{\psi\} \quad \psi \perp \omega}{\varphi \perp \omega}$$

Compositional Monotonicity (CM)	$\dfrac{S \Box\!\!\rightsquigarrow T \qquad \varphi \not\perp \psi}{S \cup \{\varphi\} \Box\!\!\rightsquigarrow T \cup \{\psi\}}$
Containment (Cm)	$\dfrac{\varphi \rightarrow \psi}{\{\varphi\} \Box\!\!\rightsquigarrow \{\psi\}}$
Context-Free Union (CFU)	$\dfrac{S \Box\!\!\rightsquigarrow T \qquad S \Box\!\!\rightsquigarrow U}{S \Box\!\!\rightsquigarrow T \cup U}$
Cut (Cu)	$\dfrac{S \cup T \Box\!\!\rightsquigarrow U \quad S \Box\!\!\rightsquigarrow T}{S \Box\!\!\rightsquigarrow U}$
Euclid (Eu)	$\dfrac{S \Box\!\!\rightsquigarrow T \quad S \Box\!\!\rightsquigarrow U}{T \Box\!\!\rightsquigarrow U}$
Guarded Left Union (GLU)	$\dfrac{S \Box\!\!\rightsquigarrow T \qquad Requirement}{S \cup U \Box\!\!\rightsquigarrow T}$
Guarded Union Containment (GUC)	$\dfrac{\varphi \rightarrow \psi \qquad S \cup \{\varphi\} \Box\!\!\rightsquigarrow T}{S \cup \{\varphi\} \cup \{\psi\} \Box\!\!\rightsquigarrow T}$
Lawyer (La)	$\dfrac{S \boxtimes\!\!\not\rightsquigarrow_{r_1} T \quad \ldots \quad S \boxtimes\!\!\not\rightsquigarrow_{r_k} T}{S \Box\!\!\rightsquigarrow_{u_2} T}$
Left Monotonic Union (LMU)	$\dfrac{S \Box\!\!\rightsquigarrow T}{S \cup U \Box\!\!\rightsquigarrow T}$
Left Related Union (LRU)	$\dfrac{S \Box\!\!\rightsquigarrow T \qquad U \Box\!\!\rightsquigarrow T}{S \cup U \Box\!\!\rightsquigarrow T}$
Left Singleton Containment (LSC)	$\dfrac{\varphi \rightarrow \psi \quad S \not\equiv S \cup \{\varphi, \psi\}}{S \cup \{\varphi\} \Box\!\!\rightsquigarrow S \cup \{\psi\}}$
Left Singleton Monotonic Union (LSMU)	$\dfrac{S \not\equiv S \cup \{\varphi\}}{S \cup \{\varphi\} \Box\!\!\rightsquigarrow S}$
Local Preclusion (LP)	$\dfrac{S \Box\!\!\rightsquigarrow \{\varphi\} \quad \varphi \perp \psi}{S \boxtimes\!\!\rightsquigarrow \{\psi\}}$

Mutual Preclusion (MP)	$\dfrac{\varphi\perp\psi}{\psi\perp\varphi}$
Negation Rationale (NR)	$\dfrac{S\boxtimes\!\!\rightsquigarrow T}{S\boxtimes\!\!\rightsquigarrow T\cup U}$
Preclusion (Pr)	$\dfrac{\varphi\perp\psi}{\{\varphi\}\boxtimes\!\!\rightsquigarrow\{\psi\}}$
Rational Compositional Monotonicity (RCM)	$\dfrac{S\ \Box\!\!\rightsquigarrow T\ \ S\equiv\{\varphi_1,..,\varphi_n\}\ \ \varphi_1\not\perp\psi..\varphi_n\not\perp\psi}{S\cup\{\psi\}\ \Box\!\!\rightsquigarrow T}$
Right Monotonic Decomposition (RMD)	$\dfrac{S\ \Box\!\!\rightsquigarrow T\cap U}{S\ \Box\!\!\rightsquigarrow T}$
Right Monotonic Relation Union (RMRU)	$\dfrac{S\ \Box\!\!\rightsquigarrow T}{S\ \Box\!\!\rightsquigarrow T\cup\{\langle\langle \mathit{Ref},\mathrm{t},\dot{\mathrm{p}};\ 1\rangle\rangle\}}$
Right Monotonic Union (RMU)	$\dfrac{S\ \Box\!\!\rightsquigarrow T}{S\ \Box\!\!\rightsquigarrow T\cup U}$
Right Singleton Containment (RSC)	$\dfrac{\varphi\rightarrow\psi\quad S\not\equiv S\cup\{\varphi,\psi\}}{S\cup\{\psi\}\ \Box\!\!\rightsquigarrow S\cup\{\varphi\}}$
Right Singleton Monotonic Union (RSMU)	$\dfrac{S\not\equiv S\cup\{\varphi\}}{S\ \Box\!\!\rightsquigarrow S\cup\{\varphi\}}$
Right Weakening (RW)	$\dfrac{S\ \Box\!\!\rightsquigarrow T\cup U}{S\ \Box\!\!\rightsquigarrow T}$
R-Right Monotonic Composition (R-RMC)	$\dfrac{S\ \Box\!\!\rightsquigarrow\{\varphi\}}{S\ \Box\!\!\rightsquigarrow\{\langle\langle R,\varphi,\psi;\ 1\rangle\rangle\}}$
Set Equivalence (SE)	$\dfrac{S\ \Box\!\!\rightsquigarrow U\quad S\equiv T}{T\ \Box\!\!\rightsquigarrow U}\qquad\dfrac{S\ \Box\!\!\rightsquigarrow T\quad T\equiv U}{S\ \Box\!\!\rightsquigarrow U}$
Simple Anti-Aboutness (SAA)	$\dfrac{S\ \Box\!\!\not\rightsquigarrow T}{S\boxtimes\!\!\rightsquigarrow T}$

Strict Composition (SC)

$$\frac{S \mathrel{\Box\!\!\rightsquigarrow} T}{S \mathrel{\Box\!\!\rightsquigarrow} T \cap S}$$

Subset Aboutness (SA)

$$\frac{S \equiv T \cup U}{S \mathrel{\Box\!\!\rightsquigarrow} T}$$

Symmetry (Sy)

$$\frac{S \mathrel{\Box\!\!\rightsquigarrow} T}{T \mathrel{\Box\!\!\rightsquigarrow} S}$$

Transitivity (Tr)

$$\frac{S \mathrel{\Box\!\!\rightsquigarrow} T \quad T \mathrel{\Box\!\!\rightsquigarrow} U}{S \mathrel{\Box\!\!\rightsquigarrow} U}$$

Typical User (TU)

$$\frac{S \mathrel{\Box\!\!\rightsquigarrow}_{r_1} T}{S \mathrel{\Box\!\!\rightsquigarrow}_{u_1} T} \quad \dots \quad \frac{S \mathrel{\Box\!\!\rightsquigarrow}_{r_k} T}{S \mathrel{\Box\!\!\rightsquigarrow}_{u_1} T}$$

Unanimous User (UU)

$$\frac{S \mathrel{\Box\!\!\rightsquigarrow}_{r_1} T \quad \dots \quad S \mathrel{\Box\!\!\rightsquigarrow}_{r_k} T}{S \mathrel{\Box\!\!\rightsquigarrow}_{u_4} T}$$

Union Containment (UC)

$$\frac{\varphi \rightarrow \psi \qquad S \cup \{\psi\} \mathrel{\Box\!\!\rightsquigarrow} T}{S \cup \{\varphi\} \mathrel{\Box\!\!\rightsquigarrow} T}$$

Samenvatting

Systemen die aan de hand van een vraagstelling relevante informatie opleveren worden information retrieval (IR) systemen genoemd. Deze systemen spelen een steeds belangrijker rol in de informatievoorziening, zeker gezien de toenemende mate waarin documenten met ongestructureerde informatie (zoals rapporten, memo's, verslagen, foto's en video's) voor nader gebruik worden opgeslagen en het toenemend gebruik van digitale bibliotheken voor dit doel. Helaas komt het maar al te vaak voor dat opgeslagen relevante informatie, indien nodig, niet meer terug te vinden is. Dit is een gevolg van het feit dat het heel lastig is om te bepalen of een document relevant is voor een gegeven vraagstelling. Het terugvinden van relevante informatie, met uitsluiting van irrelevante informatie, wordt bovendien bemoeilijkt door het feit dat informatie niet meer in één statisch informatiedomein staat opgeslagen maar, mede door de opkomst van het digitale wegennet (Internet), zich kan bevinden in diverse, over de wereld verspreide, dynamische informatiedomeinen.

De essentie van het zoeken naar relevante informatie kan als volgt omschreven worden:

> 'Op welke wijze kan men relevante informatie onderscheiden van niet-relevante informatie met betrekking tot een zekere informatiebehoefte.'

Naarmate een informatiedomein meer informatie bevat en er meer informatiedomeinen moeten worden doorzocht, wordt de rol van een IR-systeem belangrijker. Handmatige controle van het resultaat -is alle relevante informatie nu wel gevonden?- is onmogelijk geworden. Het wordt zodoende steeds belangrijker om op een verantwoorde wijze een IR-systeem, of een combinatie van meerdere IR-systemen, te selecteren.

Om te helpen bij het maken van een verantwoorde keuze wordt in dit proefschrift een theoretisch raamwerk voor IR-systemen gepresenteerd. In dit raamwerk wordt vooral gekeken naar de wijze waarop in een IR-systeem een relevantie-beslissing tot stand komt. Aan de hand van deze studie zijn we in staat kwalitatieve uitspraken te doen over de relevantie-beslissingen van verschillende IR-systemen en kunnen we op deze manier komen tot een vergelijking van hun doelmatigheid.

Als uitgangspunt geldt dat ieder IR-systeem een bepaalde methode heeft om te beslissen of een document relevant is gegeven een vraagstelling. Deze methode is afgeleid aan de hand van een model. Een IR-model is gebaseerd op de volgende drie fundamenten:

(i) *de documentrepresentatie*
voor de meeste IR-modellen is dit gewoon een verzameling representatieve trefwoorden (keywords) maar steeds vaker gebruikt men tegenwoordig meer complexe representaties die de inhoud van een document preciezer omschrijven.

(ii) *de vraagstelling*
deze wordt meestal zo samengesteld dat deze direct passend is op de documentrepresentatie van het model. In veel modellen kan een vraagstelling worden samengesteld met behulp van connectoren zoals 'en', 'of', en 'niet'.

(iii) *de matchingfunctie*
deze functie bepaalt of een documentrepresentatie relevant geacht kan worden gegeven de vraagstelling. Sommige modellen maken hierbij gebruik van opgeslagen kennis zoals die bijvoorbeeld aanwezig is in een thesaurus. Een matchingfunctie kan in plaats van relevant of niet relevant ook gradaties aangeven door middel van een *rankingproces.*

Information retrieval onderzoekers voeren vele discussies of de aanpak in model X beter is dan de aanpak in model Y. In deze discussie kiest men vaak positie aan de hand van toetsen die plaats vinden op grote, speciaal geprepareerde testcollecties (bijvoorbeeld de TREC testcollectie die meer dan 3 gigabyte aan informatie bevat). In zogenaamde recall en precision-berekeningen worden de resultaten van de toetsen omgezet in statistische waarden, die aangeven hoe doortastend en accuraat een bepaald IR-systeem is. De recallwaarde geeft aan hoeveel relevante documenten door het systeem zijn opgeleverd ten opzichte van de in het informatiedomein aanwezige relevante documenten. Precision geeft aan hoeveel opgeleverde documenten daadwerkelijk relevant zijn. Een hoge recall geeft dus aan dat het IR-systeem min of meer alles gevonden heeft wat relevant is, een hoge precision geeft aan dat alles wat door het systeem gevonden is, ook relevant is.

In dit proefschrift wordt, in plaats van een experimentele, een theoretische vergelijkingsmethode voor IR-systemen gepresenteerd. Omdat elk IR-model gebaseerd is op een geschikt begrip van 'relevantie', wordt eerst onderzocht hoe dit begrip kan worden geformaliseerd. In 1971 introduceerde Cooper een objectieve notie van relevantie genaamd 'logisch relevant'. Deze notie plaatst het begrip relevantie in een logische context, en onttrekt het aan subjectieve interpretaties. Bij logische relevantie gaat het erom of men op een logische wijze een relevantie-beslissing kan afleiden. Om verwarring tussen de begrippen 'relevant' en 'logisch relevant' te vermijden, gebruiken we de term *omtrentheid* (in het engels 'aboutness') om aan te duiden dat informatie omtrent andere informatie is. In 1986 presenteerde Van Rijsbergen het idee om te onderzoeken of er een logica, dus een taal en een formeel bewijssysteem, bestaat die de omtrentheid-relatie kan definiëren. In dit proefschrift wordt aangetoond dat dit mogelijk is. Dit is vervolgens het uitgangspunt van onze vergelijkingsmethode: stel dat omtrentheid is te karakteriseren in termen van een logica, dan kan van ieder IR-model een bewijssysteem van omtrentheid gegeven worden. Zo kunnen we dus IR-modellen aan de hand van hun bewijssystemen gaan vergelijken.

In dit proefschrift worden de omtrentheidsbeslissingen van een aantal bekende IR-modellen onderzocht en vervolgens vergeleken. Daarvoor wordt eerst in hoofdstuk 3 een theoretisch raamwerk samengesteld, waarin de fundamenten van de IR-systemen uitgedrukt kunnen worden. Binnen dit raamwerk wordt een taal geformuleerd waarin representaties van documenten en vraagstellingen beschreven kunnen worden. Deze taal is gebaseerd op de zogenaamde Situation Theory. De representaties van documenten en de vraagstellingen worden vertaald naar situaties. Rest de vraag wanneer een bepaalde situatie omtrent een andere situatie is.

Om deze vraag te beantwoorden presenteren we een aantal axioma's en afleidingsregels (tezamen postulaten genoemd). Deze postulaten drukken bepaalde karakteristieke eigenschappen van 'omtrentheid' uit. Zo is er bijvoorbeeld de regel Symmetry. Deze regel stelt dat er geen enkel verschil bestaat tussen concluderen dat situatie S omtrent situatie T is en concluderen dat situatie T omtrent situatie S is. Met behulp van een taal en een keuze uit de axioma's en de regels, kan een bewijssysteem voor omtrentheid gecreëerd worden. In dit systeem kunnen we dan stapsgewijs, gegeven een aantal feitelijkheden (de axioma's) en bepaalde regels, afleiden of een situatie omtrent een andere situatie is. Deze manier van redeneren kunnen we op IR-modellen toepassen.

In hoofdstuk 4 postuleren we zes bekende IR-modellen vanuit deze invalshoek. Na de presentatie van elk model worden de taal van situaties, de axioma's en de afleidingsregels gegeven die horen bij het model. Om aan te kunnen tonen dat het bewijssysteem ook inderdaad het IR-model representeert, worden gezondheid en volledigheid theorema's bewezen. Is een bewijssysteem gezond ten opzichte van het model dan betekent dit dat alles wat in het bewijssysteem bewezen kan worden ook inderdaad een omtrentheidsbeslissing van het model is. Volledigheid stelt het omgekeerde: alle omtrentheidsbeslissingen van het model kunnen ook bewezen worden met het voorgestelde systeem.

In hoofdstuk 5 gebruiken we de theorie om IR-systemen te vergelijken. We vergelijken IR-modellen op basis van hun bewijssystemen. Sommige systemen zijn 'bevat' in andere systemen. Een systeem A is bevat in een systeem B als iedere omtrentheidsbeslissing van A ook een omtrentheidsbeslissing van B is en als bovendien de taal van A een deelverzameling van de taal van B is. In hoofdstuk 5 definiëren we verschillende niveaus van bevat zijn, om vervolgens tot een overzicht te komen op welke wijze de zes modellen aan elkaar gerelateerd zijn.

Men kan zich nu richten op de vraag wat het voor een relevantie-beslissing van een IR-model A ten opzichte van de relevantie-beslissing van model B betekent dat het corresponderend bewijssysteem van A bevat is in het bewijssysteem van B. Het is dan mogelijk om kwalitatieve uitspraken te doen over kwantitatieve grootheden zoals recall en precision. Zo wordt in hoofdstuk 5 bewezen dat als een omtrentheidsrelatie monotoon[1] is, een uitbreiding van de documentrepresentatie (zoals het toevoegen van woorden aan de beschrijving van het een document) nooit zal leiden tot een verlaging van de recall. Bovendien kunnen we uitspraken doen over de recall-waarde, en in enkele gevallen over de precision-waarde, van de gepresenteerde modellen ten opzichte van elkaar.

In hoofdstuk 6 presenteren we drie door ons onderzochte mogelijke toepassingen van de theorie. Allereerst gebruiken we de theorie om te analyseren op welke wijze men IR-systemen met elkaar kan combineren. De aandachtspunten zijn dan welke systemen aan elkaar gekoppeld kunnen worden, en op welke wijze, en of dit inderdaad leidt tot een beter resultaat. Vervolgens geven we aan dat een ordening op bewijssystemen kan leiden tot een preferentiële ordening van documenten. Bovendien kan men, gegeven een gewenste ordening op bewijssystemen, het rankingproces van IR-systemen inspecteren. Tenslotte wordt in hoofdstuk 6 getoond op welke wijze men de meta-theorie kan toepassen als modelleringsmethode voor IR ge-oriënteerde hypermedia toepassingen.

Samenvattend, met behulp van de theorie die in dit proefschrift wordt opgebouwd, kan men analyseren op welke wijze IR-systemen besluiten dat een document relevant is gegeven een vraagstelling. Deze analyse kan men op velerlei manieren toepassen. Het is mogelijk om de beslisstappen te vergelijken, te verbeteren en te koppelen. De theorie is ook toepasbaar om andere aspecten, zoals ordening van documenten en hypermedia-toepassingen, te bestuderen.

[1] Monotoon betekent hier: als voor iedere situatie S, T en U geldt dat: als S omtrent T is dan is S verenigd met U omtrent T.

Curriculum Vitae

Theodorus Wilhelmus Charles Huibers

8 januari 1966
Geboren te Valkenswaard.

1978 – 1983
H.A.V.O. aan het Hertog-Jan College te Valkenswaard.

1983 – 1986
Atheneum B aan het Hertog-Jan College te Valkenswaard.

1986 – 1987
Militaire dienstplicht vervuld.

1987 – 1992
Studie Informatica aan de Katholieke Universiteit Nijmegen.
Afstudeerverslag: Structuuronderzoek bij een Uitgeverij.

1992 – 1996
Assistent in Opleiding aan de Vakgroep Informatica van de Universiteit Utrecht.

Index

Author index