Probabilistic Argumentation Systems Applied to Information Retrieval

THÈSE

présentée à la Faculté des sciences pour obtenir
le grade de Docteur ès sciences par

Justin Picard

Université de Neuchâtel
Institut interfacultaire d'informatique

Directeur de thèse:
Prof. Dr Jacques Savoy, Université de Neuchâtel, Suisse

Co-directeur de thèse:
Prof. Dr Jürg Kohlas, Université de Fribourg, Suisse

Experts:
Prof. Dr Hans-Peter Frei, directeur Ubilab, Suisse
Prof. Dr Jean-Pierre Müller, Université de Neuchâtel, Suisse
Prof. Dr Jian-Yun Nie, Université de Montréal, Canada

IMPRIMATUR POUR LA THÈSE

Probabilistic Argumentation Systems applied to Information Retrieval

de M. Justin Picard

UNIVERSITÉ DE NEUCHÂTEL

FACULTÉ DES SCIENCES

La Faculté des sciences de l'Université de Neuchâtel sur le rapport des membres du jury,

MM. J. Savoy (directeur de thèse), J.-P. Müller, J. Kohlas (Fribourg), P. Frei (Zürich) et J. Nie (Montréal)

autorise l'impression de la présente thèse.

Neuchâtel, le 8 mai 2000

Le doyen:

J.-P. Derendinger

Summary

In this dissertation, a new logical model of information retrieval is developed and evaluated experimentally. This model is built on a general technique for uncertain reasoning called probabilistic argumentation systems (PAS), in which propositional logic and probability theory are combined to represent and handle uncertain knowledge, both in a symbolic and in a numerical way. The logical model respects the characteristics desired for a logic of information retrieval, and interprets van Rijsbergen's Logical Uncertainty Principle in an original way. Propositional logic is shown to be a convenient way to model information retrieval, at least when associated with probability theory in the context of PAS.

As an illustration, the model is adapted to retrieval in hypertexts, and can be incorporated to any retrieval system. This specialized model for retrieval in hypertexts is then evaluated experimentally on two collections which contain citation and hypertext links: the CACM collection (3.2 MB of abstracts of scientific articles) and the TREC'8 Web Track (2.3 GB of frozen Web).

The retrieval model is also experimented in a PAS-based retrieval system, which integrates term relationships coming from different thesauri: the Wordnet thesaurus and a statistical thesaurus. Term relationships will not only be used to facilitate the matching process, as is often done with query expansion, but also to refine the initial query term weights. The Wall Street Journal collection (250 MB of news stories) will be used for these experiments.

The treatment of these sources of evidence will highlight the conclusion that treating the information retrieval problem in both a symbolic and numerical way leads to a better understanding of the mechanisms involved in the retrieval process. One important contribution of this thesis is to show that logical models of IR can be applied to collections of any size without excessive computational costs, both as tools to solve specific problems and as complete retrieval systems.

Acknowledgements

I would like to express my gratitude to my phD advisor, Prof. Jacques Savoy, for being constantly present throughout the realization of my thesis. I really appreciated the freedom he gave me to explore new research directions, and his relevant suggestions.

I would also like to thank the examining committee Prof. Jurg Kohlas, Prof. Hans-Peter Frei, Prof. Jean-Pierre Muller and Prof. Jian-Yun Nie, for advising me during my thesis. I am grateful for some very useful comments, which helped me solve certain problems which obsessed along my thesis.

Thanks to all my colleagues, for making the ambiance at work so pleasant. Special thanks to Dr. Rolf Haenni, for helping me understand the subtleties of the probabilistic argumentation systems.

Contents

Summary i

Acknowledgements iii

Notation ix

1 Introduction 1
- 1.1 Fundamental notions in IR . . . 1
  - 1.1.1 The task of an information retrieval system . . . 1
  - 1.1.2 The retrieval process . . . 2
  - 1.1.3 An example of retrieval approach: the vector-space model 4
  - 1.1.4 Evaluation and comparison of retrieval system . . . 6
  - 1.1.5 Documents . . . 6
  - 1.1.6 Using other bodies of knowledge to help retrieval . . . 7
- 1.2 The uncertainty problem . . . 9
  - 1.2.1 Document representations . . . 9
  - 1.2.2 Representation of information need . . . 9
  - 1.2.3 Matching . . . 10
- 1.3 Models of IR . . . 10
  - 1.3.1 Why do we need models of the IR process? . . . 10
  - 1.3.2 Probabilistic models . . . 11
  - 1.3.3 Logical models . . . 13
  - 1.3.4 Discussion . . . 16
- 1.4 Research summary . . . 16
  - 1.4.1 Description of the approach . . . 16
  - 1.4.2 Justification of the approach . . . 18
- 1.5 Contributions . . . 19
  - 1.5.1 The development of a new logical model based on PAS . 19
  - 1.5.2 The application of PAS to retrieval in hypertexts . . . 19
  - 1.5.3 Experiments with a PAS retrieval system and thesauri . . 20
- 1.6 Outline of the dissertation . . . 20

2 Probabilistic argumentation systems 21
- 2.1 Adjoining uncertainty to propositional logic . . . 21
- 2.2 Fundamental concepts . . . 22
  - 2.2.1 Propositional argumentation systems . . . 22
  - 2.2.2 Probabilistic argumentation systems . . . 24
- 2.3 An example related to information retrieval . . . 26

3 The PAS logical model **29**
3.1 Designing the PAS . . . 29
3.1.1 The variables of interest . . . 29
3.1.2 The body of knowledge . . . 30
3.1.3 Obtaining probabilities . . . 32
3.2 The retrieval process . . . 34
3.2.1 PAS and the logical approach . . . 34
3.2.2 Choosing the hypothesis . . . 34
3.2.3 Arguments against the hypothesis . . . 35
3.2.4 Interpreting the Logical Uncertainty Principle . . . 35
3.3 Characteristics of a logic for IR . . . 37
3.3.1 Significance . . . 37
3.3.2 Information containment . . . 37
3.3.3 Intensionality . . . 38
3.3.4 Partiality and flow of information. . . . 38
3.3.5 Uncertainty . . . 38
3.3.6 An example . . . 39
3.4 Discussion and related approaches . . . 40
3.4.1 PAS and other probabilistic logics . . . 40
3.4.2 The inference network model . . . 40
3.4.3 Probabilistic Datalog . . . 41
3.4.4 The ”possible worlds” approaches . . . 42
3.4.5 Other logical approaches . . . 42
3.5 Conclusion . . . 42

4 Retrieval in hypertext **43**
4.1 Approaches to retrieval in hypertext . . . 43
4.1.1 Classical information retrieval approaches . . . 44
4.1.2 Web-based approaches . . . 46
4.2 The hypertext retrieval model . . . 50
4.2.1 Converting hyperlinks to knowledge . . . 50
4.2.2 Enhancing document content . . . 50
4.2.3 Modifying the rank . . . 51
4.3 Extensions to the model . . . 52
4.3.1 A model for computing a personalized ”popularity” measure 52
4.3.2 A model for computing hub and authority scores . . . 52
4.4 Computing probabilities . . . 53
4.4.1 A priori support . . . 54
4.4.2 Computing link assumption probabilities . . . 54
4.4.3 Semantic link probabilities . . . 55
4.5 Experiments . . . 56
4.5.1 Learning . . . 56
4.5.2 Experiments on the CACM collection . . . 57
4.5.3 Experiments on the small Web Track . . . 57
4.6 Conclusion . . . 59

5 Retrieval with thesaurus **61**
5.1 Thesaurus and information retrieval 61
5.1.1 Manual thesaurus 61
5.1.2 Statistical thesaurus 62
5.1.3 Using thesauri for IR 63
5.1.4 Discussion 65
5.2 Determining content-bearing query terms 65
5.2.1 The Cluster Hypothesis for query terms 65
5.2.2 Determining relevant terms 67
5.2.3 Verification of the hypothesis 68
5.3 Proposed models 69
5.3.1 Basic model 69
5.3.2 Considering thesaurus information 70
5.3.3 A PAS for improving query term weights 71
5.4 Experiments 73
5.4.1 Basic model 73
5.4.2 Use of Wordnet 73
5.4.3 Statistical relationships 74
5.4.4 Determining relevant terms 75
5.5 Discussion 76

6 Conclusion **79**
6.1 Main contributions 79
6.2 Open questions 80
6.3 The future of logical approaches to IR 82

Bibliography **83**

Notation

General

d_i	Document number i
q_i	Information need number i
t_i	Term number i
N	Number of documents in the collection
df_j	Number of documents in which term t_j occurs at least once
idf_j	Inverse document frequency of term t_j
f_{ij}	Number of occurrences of term t_j in document d_i
qf_j	Number of occurrences of term t_j in the query

Probabilistic Argumentation Systems

P	Set of propositions referring to the variables of interest (PAS)
A	Set of propositions, called assumptions, denoting uncertain events or circumstances
ξ	Body of knowledge represented as a logical sentence: $\xi_1 \wedge ... \wedge \xi_k$
Σ	Body of knowledge represented as a set: $\{\xi_1, .., \xi_k\}$
X	Set of probabilities assigned to assumptions
D_i	Logical proposition referring to document d_i
Q_i	Logical proposition referring to information need q_i
C_k	Logical proposition referring to concept k
T_k	Logical proposition referring to the relevance of term t_k to the information need
a_i	Assumption used in: $D \wedge a_i \rightarrow C_i$. Also denotes an "a priori assumption in: $a_i \rightarrow D_i$ (Chapter 4), and in: $a_i \rightarrow T_i$ (Chapter 5)
b_i	Assumption used in: $C_i \wedge b_i \rightarrow D$
c_i	Assumption used in: $C_i \wedge c_i \rightarrow Q$
d_i	Assumption used in: $Q \wedge d_i \rightarrow C_i$
l_{ij}	Assumption corresponding to a hyperlink, in: $D_i \wedge l_{ij} \rightarrow D_j$
r_{ij}	Assumption corresponding to a term relationship, in: $C_i \wedge r_{ij} \rightarrow C_j$

Note: this is not an exhaustive list of all the symbols used in this thesis. Some symbols may take different meanings in different portions of the text (e.g. d_i). The letters i, j and k are alternatively used for indices. If not necessary, indices are not used, e.g. D instead of D_i. For reading commodity, assumptions are in minor letters and other propositions in capital letters.

Chapter 1

Introduction

The introduction is organized as follows: first, the fundamental notions of information retrieval (IR) will be briefly explained, in order to define a general context for this work (Section 1.1). The fundamental problem of uncertainty will then be introduced (Section 1.2). Since this thesis is about a new model of IR, Section 1.3 discusses the role that play models in the development of IR (Section 1.3). Section 1.4 presents and justifies the logical model of IR proposed in this thesis, followed by a list of the contributions of this work (Section 1.5). A short outline of this dissertation closes this introduction (Section 1.6).

1.1 Fundamental notions in IR

This section provides a short introduction to some of the fundamental notions of IR, in order to define a setting for this thesis. The neophyte reader should pay attention to the terms in bold characters, which usually appear in any text related to this subject.

1.1.1 The task of an information retrieval system

The information explosion is a secret for nobody now. The Web, one of the biggest collection of stored information, has an estimated 800 million Web pages as of February 1999 [LG99], and is increasing at the rate of a few million pages per day. There are databases of textual information concerning nearly every domain of human knowledge, but while access to information gets easier, it gets harder to extract the desired information. If we ever want this information to be useful, we are faced with the problem of developing efficient and effective techniques for finding the information we are looking for.

The information retrieval problem is raised whenever a user seeks a precise piece of information from a large amount of stored information. Consider a large **collection** of objects, which could be for example written books, newspaper articles, scientific papers, Web pages, images, sounds or videos, and a user having an information need which can presumably be answered by one or some of the objects of this collection; the task of an **information retrieval system** is to find which of these objects (documents, graphics, images, sounds), if any, will help the user answer his information need. In this thesis, objects will be documents but other forms can be used.

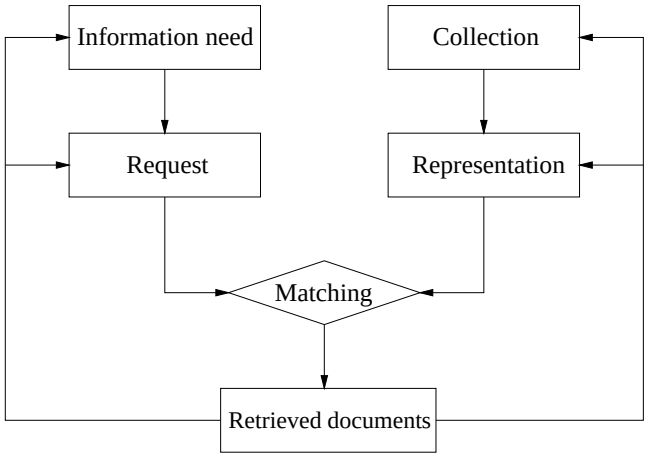

Figure 1.1: General view of the retrieval process.

1.1.2 The retrieval process

In the retrieval process, the IR system extracts the **documents** (or more generally, the pieces of information) which will presumably answer the information need formulated by the user. This retrieval process is usually separated into a preliminary step of indexing the documents, followed by an operational step of retrieval. The retrieval process can be iterative, if the user provides some feedback on the retrieval results. Figure 1.1 shows a general view of the retrieval process.

Indexing

In a preliminary step, the information retrieval (IR) system generates an internal representation of the information contained in each document, through the process of **indexing**. Usually, indexing is done by extracting **terms** from the plain text of the document. A term can designate a word, but also a stem, a noun phrase or a phrase. A stem is a word reduced to its root: for example, 'application$s$' and 'developing' would become respectively 'application' and 'develop' after stemming. The underlying (sometimes questionable) assumption behind stemming is that there is no real and important difference of meaning between the set of words which conflate to the same stem. A phrase is two or more consecutive words which have a precise signification, e.g. 'information retrieval' or 'President of the United States'. If possible, manually assigned keywords describing the contents of the document can also be used for indexing (e.g. Yahoo!). Indexing thus aims at finding the set of concepts describing each document. These concepts come often with associated weights, which represent their estimated relevance to the topic of the document. Typically, the more often a term occurs in a document and the less it appears in other documents of the collection, the better it is as a descriptor of the document. Other indicators such as the position of the keyword, the logical section from which it is extracted, or the length of the document can used to compute the weight of the related concept in the document.

In this thesis, the items designing document contents can be referred to in many ways: 'term', 'feature', 'keywords' or 'concept' amongst others. We would like to establish a distinction between what are observable items (textual or not), which can

```
Document must identify applications of fiber optic tech-
nology actually in use.

#(fiber optic) #technology #application

fiber AND optic AND (application OR technology)
```

Figure 1.2: Three examples of requests corresponding to the same information need

be referred to by 'term', 'feature', 'characteristics', 'clue', and what are unobservable items, referred to by "concept" and "information item". Sometimes, it is more appropriate to use the term 'document representative' to design both types of items without clear distinction. However the different terminologies are not equivalent, and different views of the retrieval task may underly them. In conceiving retrieval as inference, as is done in this thesis, the term "concept" seems to be more appropriate. In Chapter 3, we will come back to this discussion, when the retrieval process will be detailed.

Retrieval

The user's information need is formulated by a **request**, which constitutes the input to the retrieval system. A request can be written in natural language, as a set of keywords using a controlled vocabulary, or can be formulated with Boolean operators. The request acquisition step is an important step of the retrieval process: it can be facilitated by convenient interfaces, with the help of thesaurus containing related terms to facilitate the request formulation. Figure 1.2 shows possible formulations of the same underlying information need. The IR system gives itself an internal representation of the request, named the **query**. Query terms are often weighted in a way similar to that of the documents.

Then the IR system makes a **matching** between the query and each document representation, to estimate the degree to which it is **relevant** to the information need. This matching can be exact or soft, but due to the uncertainty inherent to the retrieval process (which will be discussed in Section 1.2), soft matching is more and more preferred, so that hard matching will not be considered in this thesis. In the case of soft matching, the documents are generally presented to the user by decreasing score, degree of match or probability of relevance. The aim of a soft matching retrieval system is of course to present the documents relevant to the information need at the top of this ranked list.

Relevance feedback

The retrieval process can be iterative, if the retrieval system receives some feedback from the user, e.g. relevance judgments of the best ranked documents. This information can be used to improve the representation of the information need, and compute a generally better ranking of documents. This process, known as **relevance feedback**, will not be considered in this thesis.

1.1.3 An example of retrieval approach: the vector-space model

The vector space paradigm to retrieval

It might be useful to illustrate how the retrieval process can be done in practice. We take here the example of a retrieval system based on the vector-space model [SWY75]. Though very simple, this model had and still has a considerable importance in the development of information retrieval, exemplified by the still widespread use of the SMART retrieval system based on this model [SM83]. In this model, retrieval has a geometric or spatial interpretation. A document d_i is represented by a point or vector in t dimensions $(d_{i1}, d_{i2}, ..., d_{it})$, where each dimension corresponds to one of the t indexing terms, and where d_{ij} is the weight given to term t_j in document d_i

[1]. In the same way, a query q can be represented by a point or vector $(q_1, q_2, ..., q_t)$. Matching a document represented by d_i to an information need represented by q can be done by measuring a similarity between their associated vectors.

Weighting and ranking schemes

One of the weaknesses of the vector-space model is that it does not provide strong theoretical arguments to support any particular weighting of terms and similarity measure. But the principle generally adopted is that the weights should be chosen in order to discriminate optimally relevant from non-relevant documents. Adopting the spatial interpretation, weights should be set such that jointly relevant documents should be as near as possible in the document space, and as far as possible from the non-relevant documents.

Query terms occurring infrequently in the documents of the collection are usually better discriminators than frequent ones, because they are more specific to the information need. The term 'computer' will not be a very good discriminator in a collection of computer science articles, where it appears many times. However, the same term can be a good discriminator in a collection such as cases law, where it appears more rarely. A common measure of term specificity is: $idf_j = \log \frac{N}{df_j}$, where df_j is the number of documents in the collection composed of N documents, in which term t_j occurs at least once.

Of course, a term occurring many times in a document is more likely to be a good descriptor of the document. Take f_{ij} as the number of occurrences of term t_j in document d_i, and (f_{ij}^{max}) as the maximum number of times that a term occurs in the document. The normalized frequency $ntf_{ij} = f_{ij}/f_{ij}^{max}$ is a measure of the degree to which a term t_j is a good descriptor of a document d_i.

These two factors influencing the discriminant effect of a term are usually combined in the so called tf-idf weighting scheme, where the weight of a term t_j in document d_i is simply the product of its ntf_{ij} and idf_j components. The weight assigned to a term is then given by:

$$d_{ij} = ntf_{ij} \cdot idf_j = \frac{f_{ij}}{f_{ij}^{max}} \cdot \log \frac{N}{df_j} \tag{1.1}$$

As is often done, query terms can be weighted using exactly the same formula. There are many variants in the way the features used here can be combined, however the tf-idf weights have been much used in IR.

[1] A list of symbols used in this thesis is given after the table of contents.

Term	df_j	idf_j	f_{1j}	f_{2j}	ntf_{1j}	ntf_{2j}	d_{1j}	d_{2j}
fiber (t_1)	10	2.30	2	2	0.4	0.5	0.92	1.15
optic (t_2)	30	3.51	3	0	0.6	0	2.11	0
application (t_3)	100	4.61	5	2	1	0.5	4.61	2.30

Table 1.1: Frequencies and weights

To measure a similarity between each document and the query, we can compute the cosine coefficient between the document vector and the query vector:

$$score(d_i, q) = \cos(d_i, q) = \frac{\sum_{j=1}^{t} d_{ij}.q_j}{\sqrt{\sum_{j=1}^{t} d_{ij}^2 \sum_{j=1}^{t} q_j^2}} \tag{1.2}$$

where t is the number of distinct terms in the collection. This measure is usually better than a projection of the document vector on the query vector, because it takes into account a measure of the length of the documents in the normalization scheme with $\sum_{j=1}^{t} d_{ij}^2$. This way, short documents are not disadvantaged.

Example

Suppose we have a collection of 1000 documents, and we wish to evaluate the degree of match of two documents d_1 and d_2 relatively to a request on 'fiber optic applications'. After stemming, the query is composed of three terms: 'fiber', 'optic' and 'application' [2]. Notice that to compute $\sum_{j=1}^{t} d_{ij} \cdot q_j$, only the query terms must be considered, since the weight of any other term is null. Assume that "fiber", "optic" and "application" are denoted by respectively the terms t_1, t_2 and t_3. Table 1.2 summarizes the characteristics of each term used to represent documents d_1 and d_2. To compute the term frequency ($ntf_{..}$) components of the weights, we assume that the maximum frequency of a term in documents d_1 and d_2 is respectively $f_1^{max} = 5$ and $f_2^{max} = 4$.

From table 1.2, we deduce that the representative vectors are
$d_1 = (0.92, 2.11, 4.61)$ and $d_2 = (1.15, 0, 2.30)$. The vector representing the information need is equal to the idf components of the query terms, because every tf value is equal to 1. The query vector is then $q = (2.30, 3.51, 4.61)$. Furthermore, we assume that documents d_1 and d_2 contain other terms such that $\sum_{j=1}^{t} d_{1j}^2 = 100$ and $\sum_{j=1}^{t} d_{2j}^2 = 50$. We may then compute the similarity between the documents and the query. For d_1:

$$sim(d_1, q) = \frac{0.92 \cdot 2.30 + 2.11 \cdot 3.51 + 4.61 \cdot 4.61}{\sqrt{100}\sqrt{2.30^2 + 3.51^2 + 4.61^2}} \simeq 0.493 \tag{1.3}$$

In the same way, we find $sim(d_2, q) \simeq 0.300$. Document d_1 would then be returned to the user before d_2.

This example is a good illustration of the techniques used to rank documents relatively to a given query, although more elaborated weighting schemes than the tf-idf are in use now. However the vector-space model has several limitations: it does not support strong theoretical arguments for choosing weights, selecting a ranking function or including thesaurus relationships between terms. In Section 1.3, we will discuss the more elaborated probabilistic and logical models.

[2] Of course, the phrase 'fiber optic' could also be considered as a term.

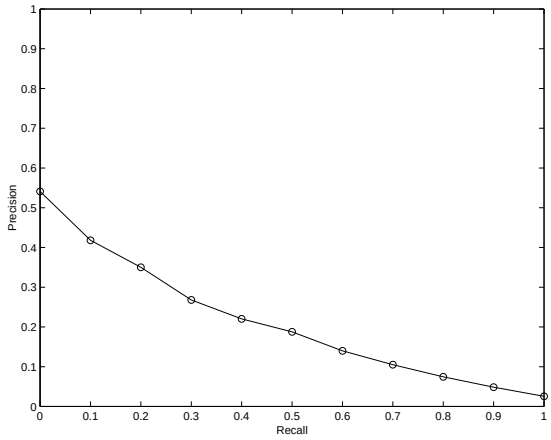

Figure 1.3: Example of a precision-recall curve obtained on 185 requests of the Wall Street Journal collection. The mean average precision at 11 recall points is 0.1995.

1.1.4 Evaluation and comparison of retrieval system

Test collections

Test collections are fundamental to IR research: they allow experimental evaluations of new retrieval models and comparisons of different retrieval systems or models. They also help tuning parameters of a retrieval system, and give directions in designing theoretical models or determining how each source of knowledge should be used to help retrieval. A test collection is made of: (1) a collection of documents, (2) a set of requests , and (3) a list of the relevant documents in the collection for each request. Often, relevance is assumed to be binary: documents are judged to be either relevant or non-relevant. Throughout this thesis, we will also assume that relevance judgments are binary.

Experimental work in this thesis will use three test collections of very different sizes: the CACM collection (3.2 Mbytes), the Wall Street Journal (246 Mbytes), and the Trec'8 Web Track (2.3 Gbytes). These collections will be described in Section 1.3.

Evaluation measure of retrieval effectiveness

It is generally accepted that the degree to which an information need has been fulfilled can be described by two quantities: precision and recall. **Precision** is the ratio (r/n) of the number r of retrieved documents which are relevant to the information need, to the total number n of documents retrieved. **Recall** is the ratio (r/R) of the number r of retrieved documents which are relevant to the total number R of relevant documents for the request in the collection. The importance respectively attributed to precision and recall varies with the user and the type of his information need. These two quantities cannot be considered independently, since recall and precision vary inversely in general.

There is no general agreement in the IR community on a universal evaluation mea-

sure of the performance of IR systems, and the choice of a more adequate measure is an open subject. Presently, the most used measure of retrieval effectiveness is the **average precision at 11-point recall**, the average on the set of requests of the precision values obtained for the 11 recall values 0.0, 0.1,..., 0.9, 1.0. This measure of retrieval effectiveness can be interpreted as the area under the precision-recall curve (see Figure 1.3). We will take this measure of retrieval effectiveness to evaluate and compare different retrieval approaches, with the use of the `TRECeval` software.

1.1.5 Documents

One of the major changes that the field of IR faced since the mid eighties concerns the notion of document. Originally, documents were considered as non-decomposable, unique and independent entities. But if structure is added to documents, the information items become decomposable units of other information items (chapters, paragraphs). Also, it gets more and more frequent to find documents referencing others, in on-line collections of documents. With the use of hypertext links, documents become interrelated and cannot be considered completely independent. An increasing proportion of the stored textual data now are hypertext documents, e.g. the Internet. One chapter of this thesis will be devoted to the use of this additional source of information to enhance retrieval.

Another major change is the nature of the information sought. Originally, information retrieval was concerned with textual data, for which an indexing based on weighted keywords may reasonably be accepted. But the concept of document has evolved from plain text to include now images, sounds, videos, graphs, tables, and others. In the most general case, a document can be a mixture of these with plain text. It can be easily understood that both conceptually and practically, representing documents with sets of weighted terms has important limitations, the most obvious being the loss of structure information. Techniques which can be accepted for IR in pure text are often too limited to represent multimedia documents. One important research topic is the development of models of IR general enough to represent these features. An example is relational indexing, in which the concepts describing the documents are allowed to have relationships [OP98].

1.1.6 Using other bodies of knowledge to help retrieval

The most important sources of evidence about document relevance are certainly the terms shared with the query. But often, documents relevant to the information need do not share enough common terms with the query. How can these document be retrieved, or at least be ranked better than they are? There are many other sources of evidence which can help retrieval if used adequately. There has been a large variety of empirical studies, in asserting whether each of these sources of evidence can serve, and how, as an additional clue for retrieval. Here are some example of these other sources of evidence:

- **Multiple information need representations.** If the user can specify his information need by more than one request (e.g. a natural language and a Boolean request), the performance obtained by combining the corresponding queries can be shown to be on average superior to each of the query [TC91, BKFS95].
- **Multiple document representations.** Different retrieval system, which make different processing of text, can capture different features from documents and

information needs. Indeed, experiments have shown that combining adequately different document representations (or results of different retrieval systems) can lead to better retrieval effectiveness than simply taking the document representation or the retrieval system which produces the best retrieval results on average [CS00].

- **Hypertext links.** Citations or hypertext links refer generally to documents dealing with related topics. When they are available, their inclusion in the retrieval process may lead to better retrieval effectiveness [FNL88, CT93, Sav94, FS95, Pic98].

- **Thesaurus.** The limitations of natural language for representing documents and matching documents and information needs, can be compensated by the use of thesaurus and other knowledge bases. A **thesaurus** contains some relationships between the set of words of a prescribed vocabulary, for example the Wordnet thesaurus used in this work contains e.g. synonymy relationships. It is also possible to build statistical thesaurus reflecting the relationships in the concerned collection. Increasing the possibilities of matching can be done by a process known as **query expansion**, where an extended query is formed after adding some terms related to the original query. A term can be added if it is related to only one query term, but thesauri are best used by exploiting the "combining evidence" paradigm: in query expansion, the contribution of each query term should be considered for choosing expansion terms [QF93, RTT99]. Moreover, the quality of the term relationships is much improved when combining the evidence from different statistical and manual thesauri [RTT99].

The results clearly suggest that IR systems can benefit from combining adequately all potential clues: multiplying the sources of information should compensate partially their fundamental uncertainty. This was stated first in the Principle of Combination [FNL88]:

> "Effective integration of more information should lead to better information retrieval."

This thesis is about the development of a new model of IR. It seems clear that a new model of IR should be able to integrate the results of previous experimental work. Surprisingly, this is not always the case: building a flexible model is a very hard problem and the most popular existing models, the probabilistic and vector-space models, are not flexible enough to allow information of very different nature to be combined, at least on the theoretical ground. The inference network model [TC91, TC92], where probabilistic dependencies between variables are combined to assess the probability that an information need is met given a document as evidence, is to our knowledge the only model which addresses precisely this issue.

1.2 The uncertainty problem

It will be argued throughout this thesis that IR should be considered as a reasoning process under uncertainty. Nearly every source of knowledge useful to the IR process is accompanied with uncertainty and imprecision. This section makes a survey of the main sources of uncertainty in IR, and of the dominant approaches to cope with this

problem. It is usual to distinguish three main sources to this uncertainty [TC97]: the document representation, the representation of the information need and their matching.

1.2.1 Document representations

At the present time, the majority of techniques for representing documents are essentially equivalent to assigning to a document a subset of a set of possible concepts $\{c_1, ..., c_t\}$, eventually with associated weights. Although this representation is very practical, and to our knowledge there is no other type of indexing which results in better retrieval effectiveness on general and large corpora, it is a somewhat rudimentary way of summarizing the content of a document. But even if one would assume that the information contained in a document can be properly represented using a set of concepts, there is in general no agreement on a unique representation of a document. There is a large part of subjectivity in assigning a concept to a document, so much that even professional indexers do not generally agree on the set of concepts describing a document. As shown by the experiments of Cleverdon [Cle84]:

> "If two experienced indexers index a given document using a given thesaurus, only 30 percent of the index terms may be common to the two sets of terms"

To the fundamental limitations of the language used for indexing, and to the subjectivity inherent to the indexing process, a third source of uncertainty comes from the fact that nowadays most document representations are now generated automatically, based on word counts in the documents. Assuming that such a thing as a "perfect" representation of documents exists, we are pretty far from it.

1.2.2 Representation of information need

The information need is a mental state internal to the user. The request is only a representation of this need, which is often unclear, confused, or badly formulated. As quoted by van Rijsbergen who cited Plato [vR96], that the user is able to formulate this information need is somewhat paradoxal: "And how will you enquire, Socrates, into that which you do not know?". Of course the user knows at least a little bit about what information he is looking for, otherwise he would not be able to formulate it at all. But still, the initial formulation of an information need is often very poor: after seeing some retrieved documents, the user can often formulate it again in a more precise and explicit way. Besides, in certain contexts such as the Internet, often the user is too lazy to make the effort to fully explain his information need, the extra work being not worth the expected result.

Of course, in the same way as for document indexing, more uncertainty is added when passing from the request to the internal representation of the retrieval system. Even if the terms are well chosen by the user, there can be more ambiguity in a request containing fewer terms than in a document which may contain a few hundreds or thousands: content-bearing terms can be detected with more statistical reliability in the latter case since they are likely to occur many times.

1.2.3 Matching

In the matching process, the retrieval system deals with two representations embedded with uncertainty, and must determine whether a concept which is poorly understood, relevance, applies to this pair of representations. Relevance has been the subject of many studies (Mizzaro reports 157 papers on this subject from 1957 to 1994 [Miz97]), but still very little is known about what makes a user decide whether a document is relevant or not. Experiments have shown that users with apparently the same background knowledge will have very different vision of what constitutes a relevant document to a given request:

> if two scientists or engineers are asked to judge the relevance of a given set of documents to a given question, the area of agreement may not exceed 60 percent." [Cle84, p. 38]

Moreover, the matching process usually assumes that term matching implies concept matching, while depending on the context, the same term may refer to different concepts and different terms may refer to similar concepts. We will address more specifically the vocabulary mismatch problem in Chapter 5.

1.3 Models of IR

The central topic of this thesis being the development of a logical model of IR, we feel it is necessary to first give motivations regarding the necessity of developing new models of IR. Then the class of models that have had the most important impact on the IR field, the probabilistic models, are introduced. The more recent logical approach to IR is then presented.

1.3.1 Why do we need models of the IR process?

Any model of IR makes underlying assumptions on (1) how documents should be represented, (2) how information needs should be represented, and (3) how these representations should be matched. For example, the vector-space model assumes that IR can be interpreted geometrically: documents and information needs can be described by vectors, which can be matched by computing a similarity between them. Although IR has a very strong empirical tradition, the importance taken by the theoretical work in modeling in a computational way the retrieval process has been increasing since the early seventies. But what are the objectives when attempting to model the IR process? Sebastiani identifies three classes of motivations [Seb98].

(1) Models are abstractions of the retrieval process, independent of the specific architecture chosen for storing the data, retrieving documents, or acquiring and processing the request. Abstraction leading to generalization, a model provides a theoretical framework for thinking the IR task. A wide variety of retrieval approaches can then be described and characterized within one model, such that the main characteristics of an approach to retrieval are more apparent and can be better understood. (2) Models provide useful guidelines for developing an operational retrieval system. For example, theoretical arguments can be used to justify that a retrieval system should be built this way rather than that way. Also, a retrieval model can be preferred because it has been shown to be more general than other existing models. The modeling task may then influence the whole IR field in a privileged direction. (3) Models can also be useful

to compare the characteristics of different retrieval approaches in a general way. As formulated by different authors [BH94, Seb98], there is a hope that models can be compared theoretically rather than experimentally. This would reduce the number of experiments and definitely eliminate options which are theoretically unjustified.

Furthermore, (4) every approach to retrieval has underlying, often implicit, assumptions. These assumptions are sometimes misleading and if so, they should be replaced by more valuable hypotheses. The intellectual effort needed to build a model leads to put a finger on the underlying assumptions, question them, and possibly replace them by a better set of assumptions. (5) And finally, a model is in most of the time based on a well-established theoretical framework. This framework often comes with a range of well known techniques, which can provide reliable tools for the task at hand. For example, probabilistic models open the way to the use of statistical tools, vector-space models to the use of geometry and matrix computation, and logical models to techniques of inference initially developed for artificial intelligence and database theory.

1.3.2 Probabilistic models

As first studied by Zipf, texts and collections of texts have recurrent statistical properties [Zip49]. The first recourse to statistical features of text for IR can be traced back to the work of Luhn in 1957 [Luh57], and Maron and Kuhns proposed in 1960 the use of probabilistic indexing [MK60]. Later on, Karen Sparck Jones demonstrated in 1972 that the inclusion of a measure of term specificity in the ranking scheme, the idf weight seen in Section 1.1.3, results in systematic improvement of retrieval effectiveness [SJ72]. Since then, a lot of work has been done on finding a proper way to integrate this statistical feature and others in order to improve the weighting of features describing documents and information needs. Probability theory is a well established and convenient framework to exploit the statistical regularities of texts, in order to assess and combine the weight of evidence given by different clues to support document retrieval.

The first attempt to describe IR in a probabilistic way is due to Maron and Kuhns in 1960 [MK60]. Already at that time, the authors were conscious that uncertainty cannot be excluded from the identification of document content. To cope with this uncertainty, they proposed that keywords should be applied to documents with a certain probability. The interest in probability theory as a possible way to model uncertainty declined for a while, but was renewed in the seventies. In 1977, Robertson proposed the Probabilistic Ranking Principle, foundations on which most probabilistic models rely [Rob77]. Under certain conditions of independence between documents, this principle guarantees optimal retrieval effectiveness if documents are ranked according to the probability that they are judged relevant, based on all available evidence.

The binary independence model [RSJ76], proposed by Robertson and Sparck Jones, guarantees optimal performance if query terms are weighted using their probability of occurrence in relevant and non-relevant documents, thus providing a rigorous theoretical framework guiding the way query terms should be weighted. However the model needs some relevance feedback data to estimate the probability that a query term occurs in a relevant document, though it can be set to a fixed constant [CH79], assumed to follow a given function or estimated from empirical data [Gre98]. Another drawback is that the model ignores the evidence provided by the number of occurrences of a term in a document, and the influence of document length. Around that time, Bookstein and Swanston developed a probabilistic model to assess the probability that a document should be assigned a given term [BS74].

Another class of probabilistic approaches is more pragmatic: unlike the previous approaches, no attempt is made to explain IR through the model, rather it is assumed that the ranking scheme follows a parameterized model whose parameters are computed on a set of training queries with statistical techniques of regression. Linear regression was applied by Fox et al. [FNL88] and polynomial functions were fitted by Fuhr and Buckley [FB89]. A potentially more adequate approach is the logistic regression model because there is no assumption of normality on the parameters to estimate, and the target variable, relevance, is usually binary [Gey94]. In this thesis, we will sometimes make use of the logistic regression to estimate probabilities.

The current trends in probabilistic approaches attempt to obtain better estimates of the probabilities [RW97, Gre98] and to get beyond the division between indexing and retrieval probabilistic models [PC98]. It is always possible to improve probabilistic models to better fit the retrieval process, but one may wonder if probabilistic models have not intrinsic limits which cannot be surpassed. An analogy can be made with the statistical approaches (hidden Markov models) used in speech recognition for more than thirty years, which have provided much insights in the statistical aspects of speech. Even if speech exhibits statistical regularities, the fundamental nature of speech is not statistical. In the same way, the fundamental nature of IR is surely not probability and statistics: by staying on this ground, there is the risk that some fundamental features of text and IR will never be captured.

But where do the limits of probabilistic approaches come from? Probabilistic models require strong independence assumptions between terms. They lack flexibility for integrating the different sources of knowledge which influence or help retrieval. Noticeable attempts have been made to relax the strong independence assumptions on which most probabilistic models are based by taking into account statistical co-occurrence between terms [vR77], but this did not lead to retrieval improvement [SvR83]. A quote of Robertson illustrates the limits of the probabilistic models [RW94]:

> "One problem with the formal model approach is that it is often very difficult to take into account the wide variety of variables that are thought or known to influence retrieval. The difficulty arises either because there is no known basis for a model containing such variables, or because any such model may simply be too complex to give a usable exact formula."

The strengths and weaknesses of probabilistic models culminate with the Okapi probabilistic model [RWHB95]. This model takes into account the most important factors which influence the weight of a term. The retrieval systems based on the Okapi approach are regularly among the best in the TREC experiments, and its reputation is now well established. However the value of a number of parameters must be estimated on a set of training queries, and there is no clear basis for justifying the values that should be taken by parameters to optimize retrieval effectiveness. Paradoxically, the attempt to move away from ad hoc retrieval has resulted in a comeback to more empirical but effective retrieval approaches.

A very different probabilistic approach is taken by Turtle and Croft, in which the retrieval process is described with inference networks [TC91]. An inference network is defined by a directed acyclic graph which represents the probabilistic dependencies between the variables. This approach considers that the probability of relevance can be computed indirectly, by computing the probability that the information need is met given a document as evidence: $p(relevance|d,q)$ is estimated by $p(q|d)$. Any probabilistic dependency can be included, as long as no cycle is created in the network.

However, this last approach can be considered as belonging to a different paradigm, in which retrieval is viewed as uncertain inference.

1.3.3 Logical models

Although the models of IR based on probability theory have a firm theoretical basis, they are still confined to the keyword-based approach: probabilistic models essentially provide theoretically sound techniques to assess the weight of evidence in favor of relevance provided by the occurrence of a query term in a document. Needless to say, the proper assessment of this weight is a crucial problem, but even with very refined techniques for assessing weights of evidence, keyword matching remains a very simplistic way of making the retrieval process. IR needs more sophisticated knowledge representation procedures, because retrieval systems must support searches in more and more heterogeneous environments, such as free text mixed with databases, spreadsheets, etc. The logical approach can be seen as an attempt to develop models which will propose a more accurate representation of documents and information need, and a more elaborated matching process than the classical approach to retrieval. Logical models can integrate naturally other bodies of knowledge as well as the too often ignored user in the retrieval process.

IR is inference

The strongest argument in favor of the logical approach to IR is that retrieval is uncertain inference, and there is probably no better way than logic to address the problem of dealing with uncertain inference:

> ”Whatever the mechanism for the treatment of uncertainty is, we strongly believe that logic should play the central role in the entire inferential approach.” [Nie96]
>
> ”This author is convinced that retrieval is inference [...]”[vR89]
>
> ”In this view, information retrieval is an inference or evidential reasoning process [...] [TC91, p. 187]).

For Sebastiani, it is essential to better understand the IR task if we ever want to significantly improve retrieval systems [Seb98]. If one considers that retrieval is inference, then logic is an adequate formalism to explore the IR task, and it should be investigated thoroughly with different logics. Logic has brought much to the development of many sub fields of computer science, especially those related to artificial intelligence. These other disciplines in computer science ”have gained much deeper insights being analyzed by means of logical techniques”, and ”a perspective from which they have been able to take advantage of razor-sharp analytical tools” [Seb98].

To understand better IR, logic is a formalism that must be explored. But how are logic and IR connected?

The connection between IR and logic

The connection between IR and logic was first underlined by van Rijsbergen in 1986, who showed that different classical retrieval approaches are essentially variations in the way of evaluating the uncertainty that the document D should imply the query Q, denoted $P(D \rightarrow Q)$, where $P(.)$ is an uncertainty measure to be defined [vR86]. In

this paradigm, which makes a link with deductive models of database theory [Seb98], a piece of information such as a document "answers" an information need if by using a sequence of logical operations, this information need can be inferred from the document. It is assumed that the terms of the implication can represent some notion of information content: a document is relevant if it explicitly or implicitly contains the information requested by the user, in which case it is possible to infer the information contained in the query from the information contained in the document. The event of relevance is not directly represented, but is implicit in the implication $D \rightarrow Q$. One basic assumption of the logical approach is that documents and queries can be modeled with logic: a query is usually seen as a logical sentence in the chosen logic, and a document as a set of logical sentences.

Nie investigated more thoroughly van Rijsbergen's initial idea of describing relevance by an inference process [Nie89]. He distinguishes two aspects of relevance, **exhaustivity** and **specificity**, whose importance depends on the user and his type of information need. He showed that for D to imply Q, D must deal with every aspects of Q: $D \rightarrow Q$ addresses the exhaustivity aspect of relevance. However a very large document (such as the Encyclopedia Universalis) filling up this condition might not be considered relevant by the user, because it is not specific enough to the information need. Specificity can be computed by reversing the order of implication ($Q \rightarrow D$): for Q to imply D, Q must deal with every aspect of D, in other words D cannot deal with a topic not present in Q. By combining the two interpretations of relevance, Nie made the demonstration that the Boolean, vector-space and probabilistic models are only specific ways to compute the uncertainty in these two implications [Nie89]. Although Nie and Brisebois showed later that specificity can also be addressed with the $D \rightarrow Q$ implication [Nie96], it is our opinion that choosing Q or D as the starting point of inference is a matter of convenience, and depends on the way the retrieval process is addressed.

From the very beginning, proponents of the logical approach have been aware that it is generally not possible to establish with certainty that a document implies a query, thus the inference process should be closely associated with an appropriate measure of uncertainty $P(D \rightarrow Q)$. Van Rijsbergen proposed a very general approach as a mean of measuring uncertainty, based on the minimal addition of information needed to establish the truth of the implication. The approach is stated in the famous Logical Uncertainty Principle [vR86, vR89]:

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

This thesis will propose an original, symbolic interpretation of this Logical Uncertainty Principle.

Logical models ten years later

Treating IR with logic has brought much enlightening on underlying mechanisms which could hardly be studied with other approaches. The logical approach has led to the creation of meta models of IR, in which different models of IR can be analyzed to illustrate if they possess some general properties. In that vein, Nie made the demonstration that some forms of vector-space model are not correct [Nie89]. Huibers and Bruza determined a set of axioms concerning information carriers, and demonstrated that Boolean

retrieval is superior to some form of coordination match [BH94]. Crestani and van Rijsbergen explored the mechanism of probability transfer in IR, using logical imaging [CvR95].

Logical models of IR are expressive. There might be no better formalism than logic to capture knowledge, and to reason on that knowledge. This expressiveness makes them a very attractive framework for knowledge-based IR [Fuh95]. They can also accommodate naturally structured documents [Lal97], and are well suited for non standard IR in multimedia environments [FGR98, OP98]. For retrieving images, Ounis obtained much better results using relational indexing than with keyword-based SMART retrieval system [OP98]. In general, they are very flexible in integrating relationships between descriptors of documents and queries, as shown in [Nie96] and in [CvR95].

At present time, we do not know yet if the whole IR field can be revolutionized by the advent of logical IR systems. It is true that implementations of large-scale information retrieval systems based on logic have been relatively rare until now. There have been attempts [CRSvR96], but it seems that logical models are not yet ready to face the realm of experiments. This has led some people to doubt about the potential utility of logic for building retrieval systems [Lal98]. But the fact that logical models have been rarely experimented has for consequence that their behaviour in practice is not well understood, and also it is not clear how the parameters describing uncertainty should be assessed.

The inference network model can be considered as an extension of probabilistic approaches, but also as a logical model with a computational flavor. Its two main motivations are (1) the need of a formal framework to combine all available evidence, and (2) the belief that retrieval is uncertain inference. Inference network can be considered as a computational model of inference, where the set of allowed inferences is predefined, and where emphasis is put on assessment of uncertainty. This is unlike other logical models, where symbolic treatment of uncertainty dominates and usually precedes its numerical assessment, and inferences do not have to follow a predefined structure. However inference network have been successfully implemented in INQUERY [CCH92], and can be regarded as a good starting point for moving away from classical retrieval approaches.

Logical models represent a promising avenue which is worth being explored. Logic has already shown to be a much enlightening framework for exploring the IR task [Nie89, BH94, CvR95], it remains now to demonstrate that logic can be the pillar of powerful information retrieval systems. And the definite proof can only be made through experiments of logical retrieval systems. This thesis, in which various experiments will be made with the logical model developed, represents a step towards this objective.

1.3.4 Discussion

The merits and weaknesses of probabilistic and logical models have been presented. It seems to us that these approaches do not result from opposite or irreconcilable views of IR, but should be considered as complementary views of the extremely complex task of giving the ability to computers to understand natural language text and filling up an information need. To some extent, these approaches can both be considered as different ways of combining all possible clues to evaluate the relevance of a document to an information need. In one case, emphasis is put on drawing inference chains between documents and the information need, in the other on the precise assessment

of the weight of evidence of each clue. Rather than excluding each other, each approach should benefit from the experimental or theoretical work done with the other approaches.

The point made here is that promising avenues in IR are to combine different views of the retrieval process. Empirical investigations should accompany closely the development of models, not follow it, as is strongly suggested in [Gre98]. The logical model developed in this thesis will be thoroughly experimented, and this will highlight some of the difficulties not always thought of when thinking the retrieval task in an abstract way.

1.4 Research summary

This section begins with a description of the work done in this thesis on developing a logical model of IR based on probabilistic argumentation systems (PAS). We also feel it is necessary to give motivations regarding the reasons why PAS are an interesting track to follow among possible logical approaches.

1.4.1 Description of the approach

The comparison of different approaches for dealing with the uncertainty underlying the retrieval process leads us to the conclusion that the most promising formalisms are those which combine coherently techniques for assessing uncertainty such as probability theory, with more flexible and accurate techniques to represent uncertain knowledge and combine the evidence from different clues for retrieval. A more precise representation of knowledge will reduce the amount of uncertainty introduced when transforming document and requests into their representative, but the loss of information and the addition of noise remain a major part of this process. Moreover, uncertainty is an important aspect when dealing with symbolic sources of knowledge such as a thesaurus or the hypertext structure. To provide an equilibrium between the symbolic and probabilistic modeling of knowledge, frameworks which separate the reasoning process from the estimation of uncertainty seem to be indicated.

The goal of this thesis is the development of an applicable logical model of IR. The model is developed within the theory of probabilistic argumentation systems (PAS), and is the first application of this technique of uncertain reasonning to IR. Probabilistic argumentation systems represent a new and unique way of integrating and synthesizing logic and probability theory into one coherent framework. The theoretical foundations of PAS are propositional logic, probability theory and the theory of evidence. Uncertain knowledge is represented in propositional logic by using a special type of proposition called assumptions, and it is assessed according to probability theory.

When a PAS is designed, it can be used to evaluate different hypotheses, given certain observations. A PAS designed to contain all necessary knowledge for retrieval in a test collection can be used, for example, to evaluate the hypothesis that the query is true given that a certain document is observed. The evaluation of a hypothesis can be done symbolically by finding all symbolic arguments supporting it, and numerically by assessing the reliability of the arguments. The central idea of PAS is then to find the arguments supporting or discarding a hypothesis of interest. Loosely speaking, an argument can be understood for the moment as ”a chain of possible events or a particular combination of circumstances that allows to deduce the truth or the falsity of the hypothesis from the given knowledge” [HKL99].

A significant part of this thesis will be devoted to the development of the logical model. The symbolic modeling of knowledge, the computation of probabilities, the way the retrieval process is done will be clearly explained. The relationships with the Logical Uncertainty Principle will be made explicit. The verification that the model respects the properties that, according to Lalmas [Lal98], should be those of a logical model, will be thoroughly investigated.

In the preceding section, it was said that logical models have not been experimented enough yet. In this thesis we will make various experiments with the logical model developed. In one part, the logical model will be adapted to treat the specific case where a ranking of document handed out by a retrieval system can be improved after integration of the hypertext links. Experiments will be done on the CACM and TREC'8 small Web track collections described below. In another part, a retrieval system entirely based on the PAS model will be experimented. Special care will be taken in integrating various term relationships from different thesauri in the retrieval process. The Wordnet thesaurus and a statistical thesaurus will be considered to design reliable term relationships. The term relationships will be used in two ways: (1) to facilitate matching in a logically-based query expansion, and (2) in a novel way where relationships between query terms will be taken into account to determine which terms are the most reliable descriptors of the information need. The logical retrieval system developed will be experimented on the Wall Street Journal collection described below. Throughout the experiments, much care will be taken to the estimation of probabilities.

Here is a description of the collections used for experiments:

- CACM (3.2 Mbytes, 50 requests): this is a small test collection on 3204 abstracts of scientific articles of the collection of the ACM. The interest of this collection relies in the citation links between the articles, which can be used to experiment with the model for retrieval in hypertexts.

- TREC'8 Web Track (2.3 Gbytes, 100 requests): this is a large collection of 247'491 Web pages, coming from 956 Web sites. This collection was designed from the TREC experiments, and is extracted from 100 Gbytes of "frozen" Web containing a high density of relevant documents. It will also be used to experiment the model for retrieval in hypertexts.

- Wall Street Journal (240 Mbytes, 185 requests): this is a medium-size collection of news stories written between 1990 and 1992. This collection was used for the TREC experiments and thus contains a large number of requests with reliable relevance judgements. This collection will be used to implement the logical retrieval system, integrating statistical and manual thesauri.

1.4.2 Justification of the approach

PAS are a completely new approach to logical information retrieval and as such, is interesting to explore. However there is a stronger, controversial reason for exploring this avenue: since the very beginning of the logical approach, there has been a strong belief that propositional logic is not suitable for IR modeling [vR86, vR89, CC92, Nie89, Lal98]. Paradoxically, the interesting features of propositional logic are at the same time recognized by those who reject it: "The difficulty is how to extend the Boolean logic without at the same time losing its advantages" [vR96]. But Sebastiani recently showed that some of the reasons for excluding Classical Logic were not really

substantiated, rather they depend on the point of view attached to the context of the implication rule. He concluded that propositional logic would be worth being explored:

> "[...] Our arguments suggest that the rebuttal of classical (propositional) logic has been too hasty. Reconsidering classical logic, maybe sieving out the real reasons of its inadequacy to the IR case, may thus be a way of better understanding the merits of non-classical approaches."

This thesis represents the first attempt to model information retrieval using propositional logic. We do agree that propositional logic $alone$ is not suitable for IR, because uncertainty cannot be computed. However, with PAS, uncertainty can be represented within propositional logic, and as will be shown, this permits the visualization of the inference processes. In this thesis, we intend to show that propositional logic can be a much more powerful IR modeling tool than commonly believed, with the additional condition that it be properly associated with a theory of uncertainty, such as probability theory.

PAS are much more than a combination of propositional logic and probability theory. More generally, PAS can be built on set constraint logic, where the possible values taken by variables is a finite number. However, binary values of variables open the way to all the techniques of inference developed for propositional logic. On another side, PAS are an instanciation of a general theory of evidence, which can be considered both on a symbolic and on a numerical ground. It can then be linked to other work done on the application of the theory of evidence to information retrieval [Lal97, LR98]. The PAS model developed here can then be extended in many ways, and is supported by solid and extended theoretical foundations. The approach taken here should be understood as one possible way of applying probabilistic argumentation systems to information retrieval. Finally PAS belong to a general theory of evidence, and applying PAS to IR can also be regarded as a first step towards application of more and more general frameworks.

Logical models would perhaps be more popular among the IR community if more people had a good knowledge of logic. Indeed, applying modal logic to IR is less intuitive than considering documents and queries as vectors, for example. However, the majority of people understand propositional logic, or at least have an intuitive idea of it. And as will be shown, the use of "assumptions" to represent uncertainty and the finding of "arguments" to evaluate a hypothesis matches the way human reason under uncertainty. This makes it an attractive framework to explore for those people unfamiliar with logic.

1.5 Contributions

This section summarizes the contributions of this thesis. The work done in this thesis can be separated in three parts, and the contributions in each part can be discussed separately:

1. The development of a new logical model based on PAS.
2. The application of PAS to retrieval in hypertext collections.
3. The experiments with a PAS-based retrieval system, which integrates different thesauri in retrieval.

The following subsections review the main contributions in each of these parts.

1.5.1 The development of a new logical model based on PAS

This part, developed in Chapter 3, is concerned with describing the theoretical aspects of the model. The main contributions of this new model are:

- A model of IR in which each decision can be explained, since the way inferences are done allows to keep track of them. Moreover, the model can handle negative evidence: arguments for but also against a given hypothesis can be found.

- The demonstration that propositional logic does not have the limitations which prevent it from modeling information retrieval. To the contrary, the modeling highlights many aspects of the retrieval process.

- It is demonstrated that the Logical Uncertainty Principle [vR86] underlies the whole PAS approach. Moreover, the principle is generalized to encompass symbolic amounts of information.

1.5.2 The application of PAS to retrieval in hypertexts

In this part developed in Chapter 4, the logical model is adapted to integrate hypertext links. An operational retrieval system taking this knowledge into account is then experimented on the CACM and Trec'8 Web'track collection. Here are the main contributions:

- Various approaches have been taken to integrate hypertext links in the retrieval process. However, for the first time it is shown that a logical treatment of this problem can be very convenient.

- The logical representation of hyperlinks leads to a better understanding of the implicit assumptions behind the use of this evidence. This understanding has an impact on the way these evidence should be assessed, combined, and spread. The "semantics" of the links can be taken into account in a more rigorous way in the computation of probability.

- The application to a large text collection (2.3 Gbytes), demonstrate that the use of logic is not problematic even on large collections. It also demonstrates that logic can also be used as a tool in existing retrieval system, thus extending the potential applications of logic in IR.

1.5.3 Experiments with a PAS retrieval system and thesauri

- A retrieval system based on logic, which competes with the state of the art systems in IR.

- The proposal and use of a "Cluster Hypothesis for query terms", which states that similar query terms according to a statistical thesaurus are more likely to be good descriptors of the information need.

1.6 Outline of the dissertation

This thesis is organized as follows. Chapter 2 will introduce the main concepts of PAS, necessary to understand their use in the present context. Emphasis will be put on understandability rather than on a detailed description of the theory. In Chapter 3, the core of the logical model will be presented: definition and meaning of the used variables, modeling of knowledge, evaluation of a document's relevance, relationships with the Logical Uncertainty Principle. A comparison with other logical approaches to IR will also be made. In Chapter 4, we will describe how our logical model can incorporate hypertext links in the retrieval process, and discuss the experimental results obtained on the CACM and Trec'8 Web track collection. Chapter 5 will develop a complete PAS retrieval system, which integrates term relationships from different thesauri in the retrieval process. Finally, Chapter 6 will review the contributions of this thesis and give directions for future work. A discussion on the potentiality offered by logic for modeling IR will conclude this thesis.

Chapter 2

Probabilistic argumentation systems

This chapter presents the necessary theory on probabilistic argumentation systems (PAS) to understand the logical model developed in the next chapter. We shall first explain the fundamental idea of the PAS approach, which is to represent uncertainty through the use of particular propositional symbols called assumptions (Section 2.1). The fundamental concepts of this technique for uncertain reasoning can then be presented, leading to propositional argumentation systems in the symbolic case, and probabilistic argumentation systems (PAS) when probabilities are assigned to assumptions (Section 2.2). We end this chapter with an example of a PAS putting in application the notions seen in this chapter (Section 2.3). This example makes the link with the following chapter by presenting an example of a simple retrieval situation modeled with PAS.

2.1 Adjoining uncertainty to propositional logic

Propositional logic is one of the simplest and most convenient ways of encoding knowledge. An apparent drawback is that pure propositional logic seems to be unsuitable for taking account of uncertainty. However, uncertainty can be handled rather easily by adjoining particular propositions called **assumptions**. In this subsection, handling uncertainty through assumptions is discussed from a general point of view.

The simplest cases of propositional knowledge are facts and simple rules. For example, let an arbitrary statement be symbolized by P_1 ("it will rain tomorrow "). A statement can be either true or false. Although P_1 only symbolizes the statement, in some circumstances P_1 denotes the fact this statement takes the true value ($P_1 = T$). Similarly, $\neg P_1$ stands for the sentence: "it will not rain tomorrow". Furthermore, for a second proposition P_2, $P_1 \rightarrow P_2$ represents a simple rule of the form "if P_1 is true, then P_2 is also true". Thus, facts and simple rules can easily be handled by propositional logic. However, facts and rules often depend on unknown conditions or circumstances and are therefore not fully reliable.

Table 2.1 shows how uncertain facts and uncertain simple rules can be handled with propositional logic. The uncertainty that fact P_1 holds can be taken into account by the rule $a_1 \rightarrow P_1$, where a_1 is an assumption. In the same way, the rule $P_1 \rightarrow P_2$ becomes $a_{12} \rightarrow (P_1 \rightarrow P_2)$, which is formally equivalent to $P_1 \wedge a_{12} \rightarrow P_2$. Assumptions

Type of knowledge	Logical symbolization	Natural language expression
A fact	P_1	"it will rain tomorrow"
A rule	$P_1 \rightarrow P_2$	P_1 implies P_2
An uncertain fact	$a_1 \rightarrow P_1$	P_1 is true if some circumstance a_1 is true
An uncertain rule	$a_{12} \rightarrow (P_1 \rightarrow P_2) \Leftrightarrow P_1 \wedge a_{12} \rightarrow P_2$	P_1 implies P_2 if some circumstances a_{12} is true

Table 2.1: Expressing uncertain facts and rules with propositional logic.

are propositions which state the unknown conditions or circumstances upon which the facts and rules depend. If an assumption is known to be true, then the fact or rule which depends on it holds. Otherwise, nothing can be deduced from this fact or rule.

More general cases of uncertain knowledge can be similarly handled. Let γ be an arbitrary well formed formula (syntactically correct expression) in propositional logic, that somehow expresses the relation between different propositions. A well formed formula is any combination of literals with the connectors $\wedge$ and $\vee$. The corresponding case where γ is not fully reliable can then be represented by $a_1 \rightarrow \gamma$. Furthermore, it may be possible to distinguish between independent circumstances. For example, $a_1 \wedge a_2 \rightarrow \gamma$ represents a situation where γ depends simultaneously on different circumstances a_1 and a_2. From the most general point of view, uncertainty is therefore captured by arbitrary propositional formulas containing assumptions.

Most applications also require a numerical assessment of uncertainty. The numerical aspect of uncertainty is obtained by assigning probabilities to assumptions. For example, if the uncertain rule $P_1 \wedge a_{12} \rightarrow P_2$ is known to hold with probability 0.6, then we set: $p(a_{12}) = 0.6$. Note that this is conceptually different from assigning a probability to the whole logical sentence $p(P_1 \rightarrow P_2) = 0.6$, as is done in other formalisms for integrating uncertainty with logic. A discussion on this subject will take place in Sections 3.2 and 3.4.

From a knowledge base composed of uncertain facts, rules or conditions, we are interested in finding which symbolic arguments support or discard a given hypothesis (symbolic evaluation). Moreover, we want to evaluate the reliability of this support by using probabilities assigned to the assumptions (numerical evaluation).

2.2 Fundamental concepts

2.2.1 Propositional argumentation systems

Argumentation systems are obtained from propositional logic by considering two disjoint sets of propositions A and P. To improve readability, capital letters will denote propositions and minor letters denote assumptions. The elements in P are the variables of interest, and the elements in A are assumptions, intended to take account of the uncertainty.

Definition 1. *Let A and P be two such disjoint sets of propositions and let ξ be a propositional sentence in the propositional language based on $A \cup P$. Then the triple (P, A, ξ) is called a* **propositional argumentation system**. *ξ is the* **knowledge base** *of the propositional argumentation system.*

ξ is sometimes given as a conjunctive set $\Sigma = \{\xi_1, ..., \xi_R\}$ of clauses ξ_i. A **clause** is a disjunction of literals, where a **literal** is a proposition or its negation. However in such cases it is always possible to use the corresponding conjunction $\xi = \xi_1 \wedge ... \wedge \xi_R$. Each clause in Σ states a fact, rule, condition, etc., which may or may not incorporate uncertainty. A propositional argumentation system handles the qualitative part of the inference process.

Propositional argumentation systems are built on propositional logic, hence they have the same syntax and semantics as propositional logic. We assume that the reader has a basic knowledge of propositional logic, otherwise for a suitable introduction to propositional logic in the context of IR, one may read [Lal98]. The assumptions play an important role in expressing uncertain information. They are used to stand for uncertain events, unknown circumstances, or possible risks and outcomes. Note that from a formal point of view, nothing distinguishes assumptions from other propositions.

Example 1. *Suppose we are interested in a set of variables $P = \{P_1, P_2\}$, with uncertainty accounted for by the set of assumptions $A = \{a_1, a_2, a_3, l_{21}\}$, and a knowledge base $\xi = \xi_1 \wedge ... \wedge \xi_4$, where : $\Sigma = \{\xi_1 : a_1 \rightarrow P_1,\ \xi_2 : a_2 \rightarrow P_2,\ \xi_3 : P_2 \wedge l_{21} \rightarrow P_1,\ \xi_4 : a_3 \rightarrow \neg P_1\}$. The triple (P, A, ξ) then constitutes a propositional argumentation system. The equivalent propositional sentence ξ is:* $(\neg a_1 \vee P_1) \wedge (\neg a_2 \vee P_2) \wedge (\neg l_{21} \vee \neg P_2 \vee P_1) \wedge (\neg a_3 \vee \neg P_1)$ [1].

When modeling knowledge in a propositional argumentation system, arguments can be inferred from the knowledge base supporting or discounting certain hypotheses. A **hypothesis** h is any well-formed logical formula with symbols in $A \cup P$, and as such, can be true or false. Let us consider hypothesis P_1 of Example 1: under what conditions is P_1 true? An **argument** in favor of a hypothesis h is a conjunction α of literals of assumptions which, if added to the given knowledge ξ, allows to deduce hypothesis h. More formally:

Definition 2. *If h is a logical sentence on the set $A \cup P$, then a conjunction α of literals of assumptions is a* **supporting argument** *for h if: $\alpha \wedge \xi \models h$, where the symbol $\models$ means that h is a logical consequence of $\alpha \wedge \xi$.*

In that case, the hypothesis is said to be **supported** by α. Similarly, α is an argument against h if $\alpha \wedge \xi \models \neg h$. Then the hypothesis is said to be **refuted** by α.

In Example 1, an argument for P_1 is a_1, because if a_1 is true then by $\xi_1 : a_1 \rightarrow P_1$, P_1 must be true. In the same way, $a_2 \wedge l_{21}$ is another argument for P_1: if a_2 is true, then by ξ_2, P_2 must be true, and if l_{21} and P_2 are true, then by ξ_3, P_1 must be true.

There is also an argument against P_1: by ξ_4, $\neg P_1$ is a logical consequence of a_3, so a_3 is an argument against the hypothesis P_1.

When testing an hypothesis, it is useful to know which arguments support or refute it.

Definition 3. *The* **quasi-support** *of h relative to the knowledge base ξ, denoted $qs(h, \xi)$, is defined as the disjunction of all minimal supporting arguments for h. α is a* **minimal argument** *if there is no $\alpha' \neq \alpha$ such that α' is also an argument for h and $\alpha \models \alpha'$:*

$$qs(h, \xi) = \vee\{\alpha_i : \xi \wedge \alpha_i \models h, \neg\exists \alpha', \alpha' \neq \alpha_i, \xi \wedge \alpha' \models h, \alpha_i \models \alpha'\} \qquad (2.1)$$

[1] As usual in Classical Logic, an implication such as $a_1 \rightarrow P_1$ can be written as the equivalent clause: $\neg a_1 \vee P_1$.

Again in Example 1, we have found that a_1 and $(a_2 \wedge l_{21})$ are arguments for P_1, so $qs(P_1, \xi) = a_1 \vee (a_2 \wedge l_{21})$. Here $a_1 \wedge l_{21}$ and $a_1 \wedge \neg l_{21}$ would also be arguments for P_1, but they are not minimal arguments because $a_1 \wedge l_{21} \models a_1$, and also $a_1 \wedge \neg l_{21} \models a_1$. The reason for using minimal arguments is to allow for a more compact and easily understandable form for the quasi-support.

The term "quasi" expresses the fact that some of the supporting arguments for h may be in contradiction with the given knowledge. An argument α is in contradiction with a knowledge base ξ if $(\alpha \wedge \xi)$ is not satisfiable, meaning that there is no possible truth values of the propositions for which the propositional sentence $(\alpha \wedge \xi)$ is true. The fact that an argument α is in contradiction with the knowledge base ξ is written: $a \wedge \xi \models \perp$ ($\perp$ represents the contradiction). $qs(\perp, \xi)$ designates the disjunction of all minimal arguments which are in contradiction with ξ.

Loosely speaking, we say that two arguments α_1 and α_2 are in contradiction with one another if $\alpha_1 \wedge \alpha_2$ is a contradictory argument, but not necessarily α_1 or α_2 alone. In our example, the supporting arguments for P_1 are in contradiction with the argument a_3 which is against P_1. So:

$$qs(\perp, \xi) = (a_1 \vee (a_2 \wedge l_{21})) \wedge a_3 = (a_1 \wedge a_3) \vee (a_2 \wedge l_{21} \wedge a_3) \tag{2.2}$$

Definition 4. *The* **support** $sp(h, \xi)$ *of* h *is defined as follows:*

$$sp(h, \xi) = qs(h, \xi) \wedge \neg qs(\perp, \xi) \tag{2.3}$$

The reason for the support is to exclude from the quasi-support those arguments which are in contradiction with the given knowledge ξ. Still in our example, the support of P_1 is:

$$\begin{aligned} sp(P_1, \xi) &= qs(P_1, \xi) \wedge \neg qs(\perp, \xi) \\ &= (a_1 \vee (a_2 \wedge l_{21})) \wedge \neg((a_1 \wedge a_3) \vee (a_2 \wedge l_{21} \wedge a_3)) \\ &= (a_1 \wedge \neg a_3) \vee (a_2 \wedge l_{21} \wedge \neg a_3) \end{aligned} \tag{2.4}$$

2.2.2 Probabilistic argumentation systems

So far, hypotheses have only been judged qualitatively. A quantitative judgement of the situation is possible if probabilities are assigned to the assumptions, e.g., $p(a_1) = x_1, p(a_2) = x_2$, etc. In this thesis we assume that the assumptions are mutually independent, e.g. $p(a_1 \wedge a_2) = p(a_1) \cdot p(a_2)$, $p(a_1 \wedge \neg a_2) = p(a_1) \cdot (1 - p(a_2))$, etc. The reader is referred to [KH96] for a discussion on this subject.

Definition 5. *Let* (P, A, ξ) *be a propositional argumentation system, and* X *be the set of probabilities assigned to the assumptions of* A*, then* (P, A, ξ, X) *is called a* **probabilistic argumentation system** *(PAS).*

With probabilistic argumentation systems, a quantitative judgment of the situation is possible once symbolic arguments are found. The degree of support is defined as the probability of the quasi-support, conditioned on the fact that the knowledge base is satisfiable (not contradictory).

Definition 6. *If h is the hypothesis of interest, then the following conditional probability is the* **degree of support** *for h [HKL99]:*

$$\begin{aligned} dsp(h,\xi) &= p(qs(h,\xi)|\neg qs(\bot,\xi)) \\ &= \frac{p(qs(h,\xi) \wedge \neg qs(\bot,\xi))}{p(\neg qs(\bot,\xi))} \\ &= \frac{p(qs(h,\xi)) - p(qs(\bot,\xi))}{1 - p(qs(\bot,\xi))} \end{aligned} \tag{2.5}$$

A more detailed computation Formally, the degree of support is the probability that the knowledge base supports the hypothesis. It is a value between 0 and 1 that represents the support or the belief that h is true in the light of the given knowledge. This measure corresponds to normalized belief in the Dempster-Shafer theory of evidence. Clearly, $dsp(h,\xi) = 1$ means that h is completely supported by the knowledge base, while $dsp(h,\xi) = 0$ means that h is not at all supported by the knowledge base. Similarly, $dsp(\neg h,\xi) = 1$ means that h is completely discarded by the knowledge base, while $dsp(\neg h,\xi) = 0$ means that h is not at all discarded by the knowledge base. It is sometimes useful to compute the **plausibility** of hypothesis h:

$$pla(h,\xi) = 1 - dsp(\neg h,\xi) \tag{2.6}$$

It can be shown that $dsp(h,\xi) + dsp(\neg h,\xi) \leq 1$. From this, it follows directly that $dsp(h,\xi) \leq pla(h,\xi)$. The plausibility represents the degree to which the hypothesis is not in contradiction with the given knowledge.

We can compute the probability of a logical formula such as $p(qs(h,\xi))$ by putting the corresponding logical sentence in disjoint form. Suppose we would like to compute the probability of $p(A \vee B)$ from $p(A)$ and $p(B)$. The equivalent disjoint form is: $A \vee (B \wedge \neg A)$ (because $A \vee (B \wedge \neg A) = (A \vee B) \wedge (A \vee \neg A) = (A \vee B)$). We have then: $p(A \vee B) = p(A \vee (B \wedge \neg A)) = p(A) + p(B \wedge \neg A)$, because A and $B \wedge \neg A$ are disjoint events. Finally, with the independence assumption stated before, $p(B \wedge \neg A) = p(B) \cdot (1 - p(A))$, and $p(A \vee B) = p(A) + p(B) \cdot (1 - p(A))$.

Example 2. *In Example 1, suppose we assign the following probabilities to the assumptions: $X = \{p(a_1) = 0.4, p(a_2) = 0.6, p(a_3) = 0.3, p(l_{21}) = 0.5\}$. We are interested in computing the degree of support for P_1 and the degree of support for $\neg P_1$. The first step is to compute the probabilities of the quasi-support for P_1, $\neg P_1$ and $\bot$.*

For P_1, we have:

$$\begin{aligned} p(qs(P_1,\xi)) &= p(a_1 \vee (a_2 \wedge l_{21})) \\ &= p(a_1 \vee (a_2 \wedge l_{21} \wedge \neg a_1)) \\ &= p(a_1) + p(a_2 \wedge l_{21} \wedge \neg a_1) \\ &= p(a_1) + p(a_2) \cdot p(l_{21}) \cdot (1 - p(a_1)) \\ &= 0.4 + 0.6 \cdot 0.3 \cdot (1 - 0.4) = 0.508 \end{aligned} \tag{2.7}$$

Similarly, we find $p(qs(\bot,\xi)) = 0.174$, and $p(qs(\neg P_1,\xi)) = p(a_3) = 0.3$. We may compute the degrees of support for P_1 and $\neg P_1$.

$$dsp(P_1,\xi) = p(qs(P_1,\xi)|\neg qs(\bot,\xi))$$

$\xi_1 : D \wedge a_1 \rightarrow C_1$	$(\neg a_1 \vee \neg D \vee C_1)$	$p(a_1) = 0.7$
$\xi_2 : D \wedge a_2 \rightarrow C_2$	$(\neg a_2 \vee \neg D \vee C_2)$	$p(a_2) = 0.8$
$\xi_3 : C_1 \wedge c_1 \rightarrow Q$	$(\neg c_1 \vee \neg C_1 \vee Q)$	$p(c_1) = 0.7$
$\xi_4 : C_3 \wedge c_3 \rightarrow Q$	$(\neg c_3 \vee \neg C_3 \vee Q)$	$p(c_3) = 0.8$
$\xi_5 : C_4 \wedge c_4 \rightarrow \neg Q$	$(\neg c_4 \vee \neg C_4 \vee \neg Q)$	$p(c_4) = 0.6$
$\xi_6 : C_2 \wedge r_{23} \rightarrow C_3$	$(\neg r_{23} \vee \neg C_2 \vee C_3)$	$p(r_{23}) = 0.6$
$\xi_7 : C_1 \wedge r_{14} \rightarrow C_4$	$(\neg r_{14} \vee \neg C_1 \vee C_4)$	$p(r_{14}) = 0.3$

Figure 2.1: The knowledge base of a PAS. The PAS describes a retrieval situation with a document, a query and four terms.

$$\begin{aligned} &= \frac{p(qs(P_1,\xi)) - p(qs(\bot,\xi))}{1 - p(qs(\bot,\xi))} \\ &= \frac{0.508 - 0.174}{1 - 0.174} \simeq 0.4044 \end{aligned} \tag{2.8}$$

In the same way, we find $dsp(\neg P_1, \xi) \simeq 0.1525$*, from which we find:* $pla(h, \xi) = 1 - 0.1525 = 0.8475$*. In conclusion, according to the knowledge base the hypothesis is supported to a degree of 0.4044, and is plausible to a degree of 0.8475. The gap between these two values is a measure of our ignorance.*

2.3 An example related to information retrieval

The two examples of Section 2.2 illustrated the fundamental concepts of the theory. In preparation for the next chapter, it seems appropriate to illustrate these same concepts within an IR context. We will use here a simple example with a single document and a query. Some semantic relationships between the concepts are included.

Consider a PAS (P, A, ξ, X) where the set of propositions representing the variables of interest is $P = \{D, Q, C_1, C_2, C_3, C_4\}$ and the set of assumptions is $A = \{a_1, a_2, c_1, c_3, c_4, r_{23}, r_{14}\}$. The body of knowledge $\Sigma = \{\xi_1, ..., \xi_7\}$ and the set of probabilities X are given on Figure 2.1.

Within parentheses we have written the clause corresponding to each rule. Here is a possible interpretation of this knowledge base within an IR context. Clauses ξ_1 and ξ_2 represent the uncertain knowledge that document D implies concept C_1 (probability of 0.7) and concept C_2 (0.8). ξ_3 and ξ_4 represent a query which is implied by concept C_1 (probability 0.7) or by concept C_3 (0.8). However, this query should not be about C_4 (0.6). Finally, ξ_6 and ξ_7 can be seen as a small thesaurus containing various semantic relationships between concepts: there is a relation from C_2 to C_3 (probability 0.6) and from C_1 to C_4 (0.3).

In retrieval based on inference, a document D is retrieved if the query Q is implied by D. One possible way to implement this idea is to add the clause $(\xi_8 : D)$ to the knowledge base ξ (D is "observed")[2] and then to evaluate the hypothesis Q. Evaluating an hypothesis means finding all supporting and all refuting arguments for this hypothesis, and then computing the numerical reliability of these arguments. The sequence of operations is to compute: $qs(Q, \xi \wedge D)$, $qs(\neg Q, \xi \wedge D)$, $qs(\bot, \xi \wedge D)$, $dsp(Q, \xi \wedge D)$,

[2]For readers familiar with the IR logical approach, the reason for having Q as the hypothesis instead of $D \rightarrow Q$ will be explained in the next section.

$dsp(\neg Q, \xi \wedge D)$. If a numerical evaluation is our main interest, computing the support $sp(Q, \xi \wedge D)$ is not necessary.

Various methods exist for finding arguments, some of which are based on resolution [KH96], but there is no room in this thesis for a lengthy discussion of these methods.

The quasi-support of Q is:

$$qs(Q, \xi) = (a_1 \wedge c_1) \vee (a_2 \wedge r_{23} \wedge c_3) \tag{2.9}$$

Similarly, the quasi-support of $\neg Q$ is:

$$qs(\neg Q, \xi) = a_1 \wedge r_{14} \wedge c_2 \tag{2.10}$$

There is a contradiction in ξ if the arguments supporting Q and $\neg Q$ are both true. Then:

$$qs(\bot, \xi) = ((a_1 \wedge c_1) \vee (a_2 \wedge C_{23} \wedge c_3)) \wedge \neg(a_1 \wedge C_{14} \wedge c_4) \tag{2.11}$$

The probability of quasi-support for Q is found by putting it in a disjoint form. The detailed computation is displayed in order to illustrate how disjoint forms can be obtained:

$$\begin{aligned} p(qs(Q,\xi)) &= p((a_1 \wedge c_1) \vee (a_2 \wedge r_{23} \wedge c_3)) && (2.12)\\ &= p((a_1 \wedge c_1) \vee ((a_2 \wedge r_{23} \wedge c_3) \wedge \neg(a_1 \wedge c_1))) \\ &= p((a_1 \wedge c_1) \vee (a_2 \wedge r_{23} \wedge c_3 \wedge \neg a_1)) \\ &= p(a_1) \cdot p(c_1) + p(a_2) \cdot p(r_{23}) \cdot p(c_3) \cdot (1 - p(a_1)) \\ &= 0.7 \cdot 0.7 + 0.8 \cdot 0.6 \cdot 0.8 \cdot (1 - 0.7) \\ &\simeq 0.6244 && (2.13) \end{aligned}$$

The analog computation for $\neg Q$ is more direct:

$$p(qs(\neg Q, \xi)) = p(a_1 \wedge r_{14} \wedge c_4) = 0.7 \cdot 0.3 \cdot 0.6 = 0.126 \tag{2.14}$$

Finally, we find for the contradiction: $p(qs(\bot, \xi)) \simeq 0.0831$.

We may then apply Equation (2.5) to compute the numerical degree of support for Q:

$$\begin{aligned} dsp(Q,\xi) &= \frac{p(qs(Q,\xi)) - p(qs(\bot,\xi))}{1 - p(qs(\bot,\xi))} && (2.15)\\ &= \frac{0.6244 - 0.0831}{1 - 0.0831} \simeq 0.5904 \end{aligned}$$

And in the same way for $\neg Q$: $dsp(\neg Q, \xi) \simeq 0.0468$. Thus we find that the plausibility is: $pla(Q, \xi) = 1 - 0.0468 = 0.9532$.

The hypothesis Q is supported by ξ to a degree of 0.5904, and is plausible to a degree of 0.9532. The gap between 0.5904 and 0.9532 is a measure of our ignorance of the situation.

Chapter 3

The PAS logical model

The elements required in order to understand the main concepts of PAS were introduced in the previous chapter. The example developed in Section 2.3 was intended to provide a general idea of the way these elements can be combined to model the retrieval process by an inference mechanism. This chapter will develop the core of the logical model proposed in this thesis. First, the PAS retrieval model must be designed. For that it is necessary to give a clear meaning to the propositions attached to documents, queries and the other necessary items. Some of the possible ways of modeling the required knowledge will then be described. The problem of estimating the required probabilities will be discussed from a general point of view (Section 3.1). We then have everything in hand to determine in which ways the matching process can be done through inference. The links between the PAS approach and van Rijsbergen's Logical Uncertainty Principle [vR86] will also be made explicit (Section 3.2). In a survey of logical models, Lalmas presented a list of properties which should be satisfied by a logic intended to model IR. We feel it is necessary to verify whether the PAS logical model satisfies these requirements (Section 3.3). Finally, the PAS logical model is discussed from a broader perspective, through a comparison with related approaches to IR (Section 3.4).

3.1 Designing the PAS

To build a PAS describing the retrieval process, we need to define the set P of variables of interest, the set A of assumptions, the knowledge base Σ and the set X of probabilities.

3.1.1 The variables of interest

The propositions associated to the main variables of interest are those referring to the documents, to the information needs and to the concepts which allow to link documents and information needs. We prefer to name these items "concepts" instead of "terms".

The proposition referring to a document d_i will be denoted D_i. The proposition referring to an information need q_j will be denoted Q_j, and the proposition referring to a given concept will be denoted C_k. If no identification is necessary, propositions D, Q and C will be used instead.

A precise meaning must be given to these propositions. Although it makes some intuitive sense to say that a document, information need or concept is true, assigning a precise, unambiguous and non-contradictory meaning to the related proposition is not so obvious. For example, a meaning that is sometimes given to D is "document d is given as evidence". Such a definition of proposition D could not hold in the PAS approach because, for example, D can be implied by some other document, in which case D is True though it is not "given as evidence". We fall on similar problems when attempting to provide a meaning to Q or C.

To provide a meaning to propositions, we believe that the user should be explicitly considered, because this is all about satisfying his information need. It seems adequate here to invoke the concept of "infon". The infons are elementary item of information individuated by a cognitive agent [Dev91]. It is not necessary to define precisely infons for the purpose of this discussion, but an infon can be thought of a property of an object or a relationship between objects. An infon can be implied or not by the situation distinguished by the agent. For a long discussion and justification of the concept of infons, the reader should refer to [Dev91]. For the application of infons and the related situation theory to IR, the reader is referred to [vRL96].

A document implies many infons, and the satisfaction of a request may be about providing to the user one infon ('Who is the president of the United States?') or many ('I want information about traveling in France'). It is then possible to provide a meaning to propositions by considering the possession (or not) by the user of certain infons, without the need of making explicit which infons are possessed. Here is the meaning assigned to propositions D_i, C_i and Q_i:

D_i: the user possesses the infons implied by document d_i.

C_i: the user possesses the infons implied by concept c_i.

Q_i: the user possesses the infons implied by the satisfaction of query q_i.

It is of course possible to consider other propositions, referring to multiple document representations (e.g. title, summary) or multiple information need representation (natural language, Boolean).

3.1.2 The body of knowledge

We show now how different types of knowledge can be modeled. Of course, the modeling suggested is not to be considered as restrictive; there are in general many ways to model any given knowledge, and here we are just pointing out some of them.

Document representation

With unambiguous descriptions of documents and information needs, one could hope to obtain perfect retrieval. But in general, it cannot be known with certainty whether or not a document is about a concept. The problems come from the symbols that are used for communicating. When an agent communicates some knowledge, this knowledge is transformed into data (e.g. natural language). The data is a set or sequence of symbols on the meaning of which the transmitter and the receiver believe to mutually agree. The data are transformed into concepts by the other cognitive agent by the information process. But the correspondence between the symbols and concepts is multifarious, and the two cognitive agents may have different ideas in mind when using a certain symbol. Besides this, the interpretation of the symbols depends on context. This context is set by the whole set of transmitted data (e.g. the document). For example, if a document is about business, the word "target" may refer to sales whereas the same word in a

book about defense will probably refer to something entirely different. This context is very difficult to extract for a retrieval system which does not understand text. It is in no way obvious which sense the word "target" will take for the retrieval system. For example, suppose that the retrieval system observes that the word "freedom" appears 5 times in a document, and from this it concludes that the document is about the concept of freedom, ($D \rightarrow C_{freedom}$). This conclusion is not completely reliable since it is only based on observations of the related word, which may refer to different concepts. We may represent the uncertainty in the observation by a proposition a, such that we have: $a \rightarrow (D \rightarrow C_{freedom})$.

In general, if D is a document and C is a concept, then we have: $a \rightarrow (D \rightarrow C)$, or equivalently $D \wedge a \rightarrow C$. The proposition a is of course an assumption in the PAS framework, and can be assigned a probability $p(a)$ according to the reliability of the observation. Its meaning is:"Concept C is implied by concept D". A probability $p(a)$ is assigned to a, which depends on the confidence of the retrieval system in this concept.

In a PAS, the uncertain knowledge that a document D_i is about concepts C_1 to C_k can be modeled by:

$$D_i \wedge a_{i1} \rightarrow C_1, ..., D_i \wedge a_{ik} \rightarrow C_k \tag{3.1}$$

There is also a "reverse" relation leading from the concepts to the documents. The fact that a concept contains the information of a document may seem less natural. Indeed a document can be about many concepts, such that it is unlikely that the concept should contain the information of the document. We may then consider that the conjunction of all concepts implies the document:

$$C_1 \wedge ... \wedge C_k \wedge b_i \rightarrow D_i \tag{3.2}$$

In practice however, it is very unlikely that the document can be inferred from the information need if it is represented by many concepts. But one may consider that a concept may englobe all the related concepts that deal with it. In that case, it is more natural to consider that, for each of the concepts included in the representation of the document, the concept of the document can be inferred with a certain probability. We may expect this probability to be proportional to the probability that the document is about the concept, but inversely proportional to the number of concepts that the document is about. Therefore, another possible way to model the relationship from concepts to documents is:

$$C_1 \wedge b_{1i} \rightarrow D_i, ..., C_k \wedge b_{ki} \rightarrow D_i \tag{3.3}$$

Finally, an intermediate approach would be to consider that certain conjunctions of concepts can imply the document, for instance: $C_1 \wedge C_2 \wedge b_{12i} \rightarrow D_i$. However it is not obvious which conjunction of concepts should imply the document.

Information need representation

Information needs are formulated by requests, which can be considered in a way similar to document (see Equations (3.1) and (3.3)). The problem of identifying concepts from terms holds: it can even be worse because in short queries, the relevant terms cannot be identified by their frequency. On the other hand, the choice of words for denoting concepts is usually done with care by the user.

It is also possible that the query be a Boolean expression such as $C_1 \wedge (C_2 \vee C_3) \wedge \neg C_4$. In that case, the Boolean query could be represented by:

$$(C_1 \wedge (C_2 \vee C_3) \wedge \neg C_4) \wedge c_1 \rightarrow Q \tag{3.4}$$

$$Q \wedge d_1 \rightarrow C_1 \wedge (C_2 \vee C_3) \wedge \neg C_4 \tag{3.5}$$

Other bodies of knowledge

Relationships between concepts can be added to our knowledge. For example, various semantic relationships between concepts originating from a manual thesaurus, statistical co-occurrence or from domain knowledge can be added. A thesaurus such as Wordnet (see Chapter 5) contains conceptual relationships, from which one would like to derive informational relationships. Such relationships can be represented by:

$$C_i \wedge r_{ij} \rightarrow C_j \tag{3.6}$$

in which the probability that the informational relationship between C_i and C_j holds is $p(r_{ij})$. We may also imagine "negative" relationships for terms with negative correlation as measured by statistical co-occurrence. These concepts tend to exclude each other in a given context. Such relationships can help in discounting an inappropriate concept when representing a query or a document:

$$C_i \wedge r_{ik} \rightarrow \neg C_k \tag{3.7}$$

Finally, knowledge bases may include more complex rules such as 'probability theory' $\wedge$ 'logic' $\rightarrow$ 'uncertain reasoning', which lead to rules with more than one antecedent, as for example:

$$C_i \wedge C_j \wedge r_{ijk} \rightarrow C_k \tag{3.8}$$

It is also possible to consider different relationships between documents expressed as $D_i \wedge l_{ij} \rightarrow D_j$, indicating the presence of a link from D_i to D_j, and having a probability $p(l_{ij})$. Such relationships will be discussed in more detail in Chapter 4.

In summary, the retrieval process is based on a PAS defined by a set of propositions $P = \{C_1, ..., C_M, D_1, ..., D_N, Q_P\}$ representing documents, concepts and information needs, and a body of knowledge $\xi = \xi_1 \wedge ... \wedge \xi_R$ where ξ_1 to ξ_R are clauses derived from rules such as the ones presented here. This body of knowledge is expressed with symbols in $A \cup P$ where A is the set of assumptions required to represent the uncertainty. Finally a set X of probabilities assigned to the assumptions allows users to evaluate the reliability of symbolic arguments found for a given hypothesis. The next subsection discusses the problem of obtaining these probabilities.

3.1.3 Obtaining probabilities

Estimating the probabilities is one of the most difficult task faced by IR [TC97]. Reasons are that (1) the target variable, relevance, is not observable in practice unless there is some feedback of the user. We may have a set of training queries with relevance judgments, as is the case in this thesis, but (2) there often may not be enough data to estimate the required data, and (3) the assumption that past queries are representative of future ones is doubtful, especially if they come from different users [Sav94]. (4) Moreover, there is no objective way to define most probabilities, because the events

are essentially subjective and cannot be observed: we may observe words, terms, keywords, but we cannot observe concepts.

With PAS, a hypothesis has a degree of support which depends on the probability of each assumption contained in its support. A set of training queries can be used to learn the values of parameters to optimize a criteria such as the average precision on the set of all requests. However the set of parameters is huge, if one considers that there are hundred thousands of documents, and that each document may imply many concepts. In such a case it is better to learn a parameterized function for a whole class of assumptions, for example the document-to-concept ($D_i \wedge a_{ij} \rightarrow C_j$) assumptions a_{ij}.

In one case (Chapter 4), we will be able to estimate probabilities by using frequency estimates. Objection (2) does not fully apply because we will have respectively 50 and 100 learning requests for the two collections. However, objection (3) applies, but there is nothing we can do about it, except minimizing the negative impact of too "optimistic" probability estimates on requests with very few relevant documents.

In another case (Chapter 5), most events will be related to the presence or absence of concepts. In that case, objection (4) applies. To compute probability estimate, we will follow the inference network approach, in the way belief estimates of concepts are provided [TC91]. The idea is that the probability or belief that a document is about a given concept depends on some observable features related to this concept, such as the frequency of the related term in the document and in the collection. Other features can be included, such as the document's length and the position of the term in the document. A second assumption is that these features may interact according to a parameterized function. After testing several values for the parameters of the function, they find that Eqn.(3.9) is a reasonable estimate for the probability of a concept given a document, on the CACM collection. By "reasonable estimate", they mean the estimate that seems to yield the highest retrieval effectiveness on average.

$$p(c|d) = 0.4 + 0.6 \cdot tf \cdot nidf \tag{3.9}$$

where $p(c|d)$ is the probability that the concept C indexed the document D. The tf and $nidf$ values are computed similarly to Section 1.1.3, and the $nidf$ is the idf normalized by $\log N$. We will make explicit use of this technique in Chapter 5. There is another kind of probability that must be estimated: it concerns the deduction that some semantic relationships between two words (for example in a thesaurus) implies that one of the related concept is about the other one. This is modeled by $C_i \wedge r_{ij} \rightarrow C_j$. In a fuzzy modal logic context applied to query expansion, Nie [Nie96] proposes a learning technique of the "relevance strength" of a term relationship. Applied to the CACM collection with 50 queries for learning, he finds for example that the relevance strength of a holonymy relationship should be approximately equal to 0.3. If there is enough learning data, this learning technique is also able to compute refined estimates for precise relationships. For example:

$$\texttt{computer} \xrightarrow{0.27} \texttt{data processor} \tag{3.10}$$

Probabilities of assumptions can be assessed in a similar way. Chapters 4 and 5 will show how these probabilities are obtained in a real application.

3.2 The retrieval process

We are now ready to define the retrieval process. A reminder on the logical approach, already presented in Chapter 1, is necessary, to formulate it again in the context of the PAS model (Section 3.3.1). A PAS can be used to evaluate hypotheses of interest. To evaluate the relevance of a document to a given information need, a proper hypothesis must be determined (Section 3.3.2). The possible case where there are arguments against the hypothesis will then be addressed (Section 3.3.3). Finally the relationships between PAS and the Logical Uncertainty Principle will be investigated (Section 3.3.4).

3.2.1 PAS and the logical approach

The fundamental hypotheses of the logical approach to IR can be summarized in three points [CC92]. For each of these points, different interpretations are possible. We present here a convenient interpretation in the context of PAS:

1. In order to be relevant to an information need Q, a document D must logically imply Q, which is expressed by: $D \rightarrow Q$. It may also be useful to consider that the query Q must imply the document D, which is expressed by: $Q \rightarrow D$.

2. Since information and knowledge is by nature uncertain in IR, the truth of the implication cannot be established with certainty, and it is only possible to measure a degree of certainty $P(D \rightarrow Q)$ or $P(Q \rightarrow D)$.

3. This degree of certainty is evaluated through the bias of logic, following a general Logical Uncertainty Principle which in this case can be enunciated as follows:

   Given a query Q and a document D; a measure of the uncertainty of $D \rightarrow Q$ $(Q \rightarrow D)$ relative to a given body of knowledge ξ is determined by the minimal extent to which we have to add information to ξ, to establish the truth of $D \rightarrow Q$ $(Q \rightarrow D)$.

3.2.2 Choosing the hypothesis

In the PAS framework, we need to choose a hypothesis which will correspond to the $D \rightarrow Q$ or $Q \rightarrow D$ interpretation of relevance. We think that the two interpretations of relevance are useful, so the following discussion here applies to both of them. The problem is determine what hypothesis exactly will be evaluated if the $D \rightarrow Q$ or the $Q \rightarrow D$ interpretation is chosen. The following discussion considers the $D \rightarrow Q$ interpretation, but it applies as well to the $Q \rightarrow D$ interpretation. There are two possible choices:

- (1): The hypothesis is $D \rightarrow Q$, such that $(d)sp(D \rightarrow Q, \xi)$ must be computed.
- (2): The hypothesis is Q and D is added to the knowledge base (set to True), such that $(d)sp(Q, \xi \wedge D)$ must be computed.

The two cases are apparently very similar: using the fact that $\xi \wedge \alpha \models h$ is equivalent to: $\xi \wedge \neg h \models \neg\alpha$, it can be verified easily that in the two cases, the quasi-support is the same. In case (1), we have: $\xi \wedge qs(D \rightarrow Q, \xi) \models D \rightarrow Q$, which is equivalent to: $\xi \wedge D \wedge \neg Q \models \neg qs(D \rightarrow Q, \xi)$. In case (2), $\xi \wedge qs(D \rightarrow Q, \xi) \wedge D \models Q$ is also equivalent to $\xi \wedge D \wedge \neg Q \models \neg qs(D \rightarrow Q, \xi)$.

Interpretation	Measure	For	Against
$D \rightarrow Q$	Symbolic	$sp(Q, \xi \wedge D)$	$db(Q, \xi \wedge D)$
$Q \rightarrow D$	Symbolic	$sp(D, \xi \wedge Q)$	$db(D, \xi \wedge Q)$
$D \rightarrow Q$	Numeric	$dsp(Q, \xi \wedge D)$	$ddb(Q, \xi \wedge D)$
$Q \rightarrow D$	Numeric	$dsp(D, \xi \wedge Q)$	$ddb(D, \xi \wedge Q)$

Table 3.1: Summary of the possible symbolic and numerical measures of relevance.

However the support may not be the same, because the quasi-support of the contradiction, $qs(\bot, \xi)$ in case (1) and $qs(\bot, \xi \wedge D)$ in case (2), may not be the same. Any argument against D, e.g. $a_1 \rightarrow D$, will be in the quasi-support of the contradiction in case (2), but not in case (1). Moreover, in case (1) a_1 will be an argument for hypothesis.

Consequently, case (1) will lead to pathological cases if there are arguments in the knowledge base against D. Since both hypotheses are generally equivalent and we want to prevent any erroneous case, we propose $sp(Q, \xi \wedge D)$ and $sp(D, \xi \wedge Q)$ as two symbolic measures of uncertainty in $D \rightarrow Q$ and $Q \rightarrow D$ respectively. The two associated numerical measures are $dsp(Q, \xi \wedge D)$ and $dsp(D, \xi \wedge Q)$ respectively.

3.2.3 Arguments against the hypothesis

The example in Section 2.3 has shown that it is also possible to find arguments against the hypothesis Q (and similarly for D). That is, one can compute $sp(\neg Q, \xi \wedge D)$ and $dsp(\neg Q, \xi \wedge D)$. Support for $\neg Q$ expresses the notion that there may be arguments discarding the hypothesis Q, or at least re asons for doubting it. This decreases $pla(Q, \xi \wedge D)$, the degree of plausibility of Q, which is equal to 1 when there is no negative argument.

If a document is not retrieved or has very low ranking, it is probably not necessary to find any more reason to discard it. But often a large part of the documents retrieved is not relevant and is a source of noise. Finding negative arguments can be seen as a procedure for enhancing precision. This can be useful especially to those users more interested in precision than recall.

Table 3.1 provides a summary of the different qualitative and quantitative measures that can be used to evaluate the relevance of a document to an information need.

3.2.4 Interpreting the Logical Uncertainty Principle

The idea of adding a minimal amount of information is fundamental to the logical approach and it might be interesting to see how much support this idea creates in the framework of PAS. Adding a "minimal amount of information" means adding only the information essential for allowing, for example, Q to be implied by D, and nothing else. To see how this idea is respected in the PAS framework, consider the following case where a PAS is formed from: $\Sigma = \{a_1 \rightarrow P_1, a_2 \rightarrow P_1, a_3 \rightarrow P_2\}$. P_1 is a logical consequence from this knowledge base if, amongst others, one of the following is true: a_1, $a_1 \vee a_2$, a_2, $a_1 \wedge a_3$, $a_1 \wedge a_2 \wedge a_3$. However a_3 is irrelevant to P_1 and the knowledge that it is true is not necessary. More generally, some of these logical sentences are

logically weaker and more probable than others. For example, $a_1 \wedge a_2 \wedge a_3 \models a_1 \wedge a_3$, and $p(a_1 \wedge a_2 \wedge a_3) \leq p(a_1 \wedge a_3)$.

So which of these logical sentences would be the support of P_1? According to Eqn. (2.1), the support of P_1 would clearly be: $sp(P_1, \xi) = a_1 \vee a_2$. Indeed the support appears to be the most probable logical sentence which, when added to the knowledge base ξ, allows to deduce the hypothesis. And it is generally accepted that probability and information are inversely related [Nau70]: the more a sentence is a priori probable, the less learning that it is true will reduce our uncertainty. The following theorem demonstrates that the support is indeed the minimal amount of information which must be added to ξ, to establish the truth of hypothesis h.

Theorem 1. *The symbolic support of $sp(h, \xi)$ of a hypothesis h represents the minimal amount of information that must be added to ξ, sufficient to prove the hypothesis h.*

Proof. Let s be a logical sentence, and $p(s)$ be the probability that s is true. Let $Inf(s)$ be a measure of the information contained in the knowledge that s is true. $Inf(s)$ is sometimes defined as $1 - p(s)$ or as $-\log p(s)$ [Nau70] but more generally, assume that Inf is a monotonically decreasing function of p.

For the purpose of this demonstration, it is more convenient to consider sets of maximal arguments, i.e. conjunctions of literals of all assumptions. Now let $A = \{a_1, ..., a_m\}$ be the set of assumptions, and let $N_A = \{c_1, ..., c_{2^m}\}$ be the set composed of the 2^m maximal conjunctions based on the literals of A. For any hypothesis h, the set N_A can always be divided into the disjoint sets N_h of those conjunctions which support h, and $N_{\bar{h}}$ of those conjunctions which do not support h.

Obviously, the set $N_{\bar{h}}$ is of no use to hypothesis h, and the information that must be added to ξ can only be taken in N_h, meaning that any subset of $N_s \subset N_h$ supports the hypothesis. Take $N_{\bar{s}}$ such that: $N_s \cap N_{\bar{s}} = \oslash$, $N_s \cup N_{\bar{s}} = N_h$. We have:

$$\begin{aligned} p(N_h) &= p(N_s \cup N_{\bar{s}}) \\ &= p(N_s) + p(N_{\bar{s}}) \\ &\geq p(N_s) \end{aligned}$$

Since Inf is a monotonically decreasing function of p:

$$p(N_h) \geq p(N_s) \Longrightarrow Inf(N_h) \leq Inf(N_s) \tag{3.11}$$

□

The support is thus a symbolic representation of the minimal amount of information that must be added to the knowledge base. It is remarkable that this qualitative measure is independent of the probabilities given to the assumptions. The numerical equivalent of this symbolic measure of uncertainty can be evaluated, but the numerical aspect comes after the symbolic aspect. This allows the complete transparency of the inference process, the result of which is a symbolic description of the minimal information that must be added. Interestingly, this qualitative measure is expressed in the language of propositional logic, which until now has not been very popular in logical approaches.

3.3 Characteristics of a logic for IR

In presenting an introduction to logical models, Lalmas [Lal98] makes a list of the desirable properties that a logic applied to IR should possess: significance, information containment, intentionality, partiality, flow of information and uncertainty. First we will see how the PAS logical model satisfies each of these characteristics. After that, we will see how some examples of queries proposed by Lalmas to illustrate how the deficiencies of Classical Logic can be processed within the PAS framework.

3.3.1 Significance

If several items of information may represent a document or a query, they would obviously not have equal significance. A logic process must be able to represent the relative importance of each item of information. Evidently, Classical Logic alone cannot represent this requirement, because propositions can only take binary values. A possible way to represent significance is to make a quantitative assessment of the uncertainty attached to each information item; for example, with the use of the probability theory. In PAS, assumptions represent the uncertainty attached to a proposition, e.g. $D_i \wedge a_{ij} \rightarrow C_j$. The probability $p(a_{ij})$ attached to the assumption allows us to numerically represent the significance of concept C_j for document D_i.

3.3.2 Information containment

Information items are in general not independent entities, they contain information about each other. For example there is some overlap between the concepts 'uncertain reasoning' and 'probability theory', so one is about the other in an uncertain way. This property of information should be captured by the implication [Lal98, vR89]: P_1 implies P_2 if P_1 is about P_2, or if from the information in P_1 one can infer the information in P_2. In Classical Logic $P_1 \rightarrow P_2$ is equivalent to $\neg P_1 \vee P_2$, thus if P_1 is always false, the rule can imply anything, or if P_2 is always true anything can imply it: apparently, Classical Logic does not capture information containment [Lal98]. This has for consequence that a "false document" can imply any information need [CC92] . However, Sebastiani [Seb98] shows that if validity or logical consequentiality is adopted as the logical status of the implication, such problems will not occur.

In the PAS framework, a connection between two information items is denoted $P_1 \wedge a_{12} \rightarrow P_2$, where a_{12} represents the uncertainty attached to the underlying connection, which is measured by $p(a_{12})$. Obviously, $p(a_{12})$ tends to 1 when the connection is strong, and tends to 0 when it is weak. Two intuitively non-connected sentences P_1 and P_2 would have a probability $p(l_{12})$ of 0, which is equivalent to having no connection. Thus there will be no supporting argument for P_2 "coming" from P_1. In the opposite case where $p(l_{12}) = 1$, the rule becomes: $P_1 \rightarrow P_2$. In that case, the arguments for P_1 are arguments for P_2, at least if they are not arguments against P_2. There is no "false document problem", since if P_1 is false, no argument for P_2 may "come" from P_1.

Another interesting property of PAS is that it captures evidence that two information items tend to exclude each other, which is denoted: $P_1 \wedge a_{12} \rightarrow \neg P_2$. Typically, this will add arguments that lead to doubt about hypothesis P_2, since it is "incompatible" with P_1.

3.3.3 Intensionality

The meaning of textual data depends on the context. For example, in a given context C_2 will be a synonym of C_1 and in others C_3 will be synonym of C_1. It is possible to model this situation as: $C_1 \wedge r_{12} \rightarrow C_2, C_1 \wedge r_{13} \rightarrow C_3$. If it is believed that C_1 has at least one synonym, one can add the constraint: $r_{12} \vee r_{13}$. But C_2 and C_3 may refer to two different meanings of C_1 (e.g. the word "bank" in "river bank" or "Swiss bank") which are sometimes not compatible.

Various situations may be represented. For example, the case where C_1 has no more than one synonym can be expressed as: $\neg(r_{12} \wedge \neg r_{13})$. In the case where C_1 has exactly one synonym, one and only one assumption must be true. This is represented by: $(r_{12} \wedge \neg r_{13}) \vee (\neg r_{12} \wedge r_{13})$. Finally, if it is a too strong condition to have exclusive meanings, it is also possible to temper it with assumptions in the following way: $a_1 \rightarrow (\neg r_{12} \vee \neg r_{13})$.

Intentionality means that the semantics attached to a concept is dependent on the context. This context can be defined by the surrounding concepts, which can be used to capture the semantics: if the other concepts of a query refer to 'money', 'interest rates', it serves as strong evidence for the linking of the meaning of 'bank' to 'financial institution', and as evidence against the meaning 'shore of a river'.

Suppose a query contains information items C_1 and C_2. C_1 has two synonyms C_3 and C_4 ($C_1 \wedge r_{13} \rightarrow C_3, C_1 \wedge r_{14} \rightarrow C_4$) which are generally used in different contexts. In this query, the context is represented by C_2 only. If C_2 is related in any way to C_3 ($C_2 \wedge r_{23} \rightarrow C_3$), this will emphasize that C_3 is a more appropriate synonym for C_1. On the other hand, if C_2 and C_4 tend to exclude each other, as measured by a very weak statistical co-occurrence ($C_2 \wedge r_{24} \rightarrow \neg C_4$), this will be negative evidence for C_4.

3.3.4 Partiality and flow of information.

The representation of information is in general partial, whereby some information items are not originally identified, especially in written requests. As mentionned in the introduction, natural languages tend to include various implicit information. Revisions or additions of knowledge can lead to contradictions, and such cases can be easily handled by PAS. For example, if one has some argument in favor of P_1 from $a_1 \rightarrow P_1$ and later has an argument against P_1 from $a_2 \rightarrow \neg P_1$, then the argument $a_1 \wedge a_2$, which leads to a contradiction, is added to the quasi-support of $\perp$, and is considered in the computation of the support (see Eqn. (2.3). In general, it is always possible to combine the original knowledge base ξ_0 with new sources of knowledge $\xi_1, ..., \xi_N$, where the new knowledge base is to be seen as a conjunction of the added knowledge bases: $\xi' = \xi_0 \wedge ... \wedge \xi_N$. The contradictory interpretations of the assumptions are then taken into account by computing the new values of $qs(\perp, \xi')$ and $p(qs(\perp, \xi'))$.

3.3.5 Uncertainty

Uncertainty is omnipresent in IR, and is related to all the preceding properties. A numerical representation of uncertainty is a necessary feature of an IR model. But often in problems that deal with uncertain reasoning, and especially in IR, it is preferable to distinguish the symbolic and numerical aspects of uncertainty. PAS is an efficient tool for representing uncertainty in both situations.

For example, human knowledge is first expressed in symbolic form, such as domain knowledge or a relational thesaurus. In a first step, a human defined rule such as C_1 AND $C_2 \rightarrow C_3$ can be modeled by $C_1 \wedge C_2 \wedge a \rightarrow C_3$, where $p(a)$ will have to be assessed. This probability is not a fixed value, since it depends on the context (collection, user, query), and the strength of the connection may vary.

3.3.6 An example

Lalmas [Lal98] illustrates some of the issues involved in dealing with logical models by discussing different examples of queries. The documents and queries are represented in propositional logic as follows: $D = C_1 \wedge C_2$, $Q_1 = C_1, Q_2 = C_3, Q_3 = C_1 \wedge C_3, Q_4 = C_1 \vee C_3$ and $Q_5 = C_1 \wedge C_2$. These queries have been built to illustrate the deficiencies of Classical Logic. With Classical logic, D would be retrieved for Q_1, Q_4 and Q_5: indeed, one can see that $D \rightarrow Q_1$, $D \rightarrow Q_4$ and $D \rightarrow Q_5$. However the retrieval system would not make any difference between Q_4 and Q_5 while clearly D is more relevant to Q_5 which is the same as D. Moreover, D answers partially Q_3 (both are about C_1) but partiality cannot be taken into account.

A closer look at these queries reveals that D would best answer these queries in the following order: Q_5, Q_1,(Q_4 and Q_3), Q_2. Clearly, D is more likely to be relevant to Q_5 since they are identical. Q_1 comes after because D answers Q_1 exhaustively, but is not completely specific. Q_3 and Q_4 both deal with C_1 and C_3, but it is harder for a document to imply Q_3 because both concepts are necessary; while one being about Q_4 is sufficient. On the other hand, Q_3 is more firmly about C_1 than Q_4 (which can be about C_1 *or* C_3), so D appears to be more specific about Q_3 than Q_4. In summary, the ranking of Q_3 and Q_4 depends on the importance attributed to exhaustivity and specificity. Finally, Q_2 comes last because it is not in any way related to D.

With PAS, the purpose is not to establish or not relevance, but to find what arguments, if any, support or discard relevance. Following Section 3.1, knowledge can be modeled in the PAS framework in the following way:

$D \wedge a_1 \rightarrow C_1$	$D \wedge a_2 \rightarrow C_2$
$C_1 \wedge b_1 \rightarrow D$	$C_2 \wedge b_2 \rightarrow D$
$C_1 \wedge c_1 \rightarrow Q_1$	$Q_1 \wedge d_1 \rightarrow C_1$
$C_3 \wedge c_2 \rightarrow Q_2$	$Q_2 \wedge d_2 \rightarrow C_3$
$C_1 \wedge C_3 \wedge c_3 \rightarrow Q_3$	$Q_3 \wedge d_3 \rightarrow C_1 \wedge C_3$
$C_1 \vee C_3 \wedge c_4 \rightarrow Q_4$	$Q_4 \wedge d_4 \rightarrow C_1 \vee C_3$
$C_1 \wedge C_2 \wedge c_5 \rightarrow Q_5$	$Q_5 \wedge d_5 \rightarrow C_1 \wedge C_3$

In Table 3.3.6, the two symbolic measures of relevance for each query are shown. In our example, the reliability of the arguments is not assessed, but each query can be ranked according to the total number of arguments derived for the two measures $sp(Q, \xi \wedge D)$ and $sp(D, \xi \wedge Q)$. The ranking is: Q_5, Q_1, (Q_4 and Q_3), Q_2. It is remarkable that a ranking of queries is possible even though no quantitative measure is used, and that it corresponds in an overall manner to the intuitive ranking one would expect for these queries. The ranking of Q_3 and Q_4 depends on the relative importance accorded by the user to specificity and exhaustivity, which may lead to different values of assumption probabilities.

Query	$sp(Q, \xi \wedge D)$	$sp(D, \xi \wedge Q)$	# of arguments
Q_1	$a_1 \wedge c_1$	$d_1 \wedge b_1$	2
Q_2	F	F	0
Q_3	F	$d_3 \wedge b_1$	1
Q_4	$a_1 \wedge c_4$	F	1
Q_5	$a_1 \wedge a_2 \wedge c_5$	$(d_5 \wedge b_1) \vee (d_5 \wedge b_2)$	3

Table 3.2: Supporting arguments for the relevance of D

3.4 Discussion and related approaches

In this section, PAS are discussed from a general point of view. Then the pas model is compared with other logical models of IR.

3.4.1 PAS and other probabilistic logics

One may wonder if there is a fundamental difference between probabilistic logic, where probabilities are assigned to logical sentences (e.g. $p(P_1 \rightarrow P_2) = 0.6$), and PAS where probabilities are assigned to assumptions which condition events or rules (e.g. $P_1 \wedge a_{12} \rightarrow P_2$, $p(a_{12}) = 0.6$). Indeed there is a major conceptual difference: the foundations of PAS are not propositional logic but the theory of evidence. Assumption Truth Maintenance Systems (ATMS) were proposed in 1986 by de Kleer as a way to integrate uncertainty within propositional logic using assumptions [dK86]. Later, Laskey and Lehner made a step toward the integration of symbolic and numeric approaches to uncertainty management, by demonstrating that there is an equivalence between Dempster-Shafer evidence theory and ATMS [LL89]. This work was extended by Kohlas and Haenni, who established the foundations of PAS and assumption-based reasoning [KH96].

Moreover, PAS built on propositional logic are a special case where assumptions and other variables are binary. More generally, PAS can be built on Finite Set Constraint logic [HL98], where variables and assumptions can take any of a finite set of values. In an IR context, this can be useful for example to assign more than two possible values to relevance. Besides, there is a whole body of research devoted to optimize the inference algorithms for different cases, and to obtain good approximations of degrees of support. There is also an implementation of PAS in ABEL, a new modeling language for problems in the domain of assumption-based reasoning [AHKL97] In the most general case, ABEL can combine binary, multi-valued and real variables[1].

When choosing a technique for reasoning under uncertainty, it is important to verify if the inferences are correct according to human reasoning. PAS, by keeping track of inferences (which are "described" in the arguments), overcomes some of the problem of extensional semantics. For example, correlated causes do not combine in the same way as independent causes. Also, the use of assumptions allows to modify conclusions as new knowledge is added.

3.4.2 The inference network model

In the inference network model [TC91], relationships between documents, concepts and information needs are modeled by a Bayesian network, and the retrieval process is

[1] ABEL can be downloaded at: http://www2-iiuf.unifr.ch/tcs/ABEL

done by computing the probability that the query concept is met given that some variables are observed. The major difference between PAS and Bayesian networks is that Bayesian networks have a fixed hierarchical structure, allowing better computational efficiency but reducing the set of allowed inferences. Also, Bayesian nets as described in [TC91] do not allow cycles, so thesaurus or hypertext links cannot be integrated in the natural way of creating links between the variables in the net.

The representation of dependencies with Bayesian networks can be more economical than with PAS. For example, if binary variables c_1 to c_n are parents of variable d, then a link matrix of size $2 * 2^n$ is necessary in theory for describing the probabilistic dependency between d and its n parents. However, with canonical matrix such as weighted-AND [TC91], one needs only specify n parameters to define the link matrix. With PAS, one would need the 2^n rules: $C_1 \wedge ... \wedge C_n \wedge a_1 \rightarrow D, C_1 \wedge ... \wedge \neg C_n \wedge a_2 \rightarrow D, \neg C_1 \wedge ... \wedge \neg C_n \wedge a_{2^n} \rightarrow D$, plus the 2^n rules for $\neg D$. But even that would not be enough: backward relationships are not implicit in those rules, while they are with Bayesian networks.

It appears that Bayesian networks are more convenient if emphasis is put on the quantitative aspect of inferences, if one accepts the lack of flexibility of the inferences allowed (e.g. fixed structure, no cycle).

3.4.3 Probabilistic Datalog

A well established logical model is that of Probabilistic Datalog, also called Datalog$_P$ [Fuh95, FGR98, RF98, Fuh00]. Datalog$_P$ is based on first-order logic, which has been shown to be a powerful representation scheme for IR. For example, temporal relationships useful for video retrieval or spatial relationships useful for image retrieval are naturally represented within this framework [RF98]. And some experiments on an image test collection by Ounis [OP98] have shown that relational indexing leads to better retrieval effectiveness than keyword-based indexing. It appears that in some cases, Datalog$_P$ with independence assumptions and PAS make the same numerical inferences. Recursive rules are allowed in both approaches, although special caution must be taken in Datalog to avoid infinite cycles, while with PAS no special care need be taken. Also, Datalog$_P$ is limited to Horn clauses, while PAS can deal with any propositional sentence. But a point-to-point comparison of PAS with other logical models is not very interesting.

More interesting is the question of the choice of an appropriate logic for IR: what makes a logic a better choice than another ? Three possible criteria are: (1) the ability to represent knowledge, (2) the computational cost of inferences, and (3) the insight brought by the logic in the IR process. Obviously, first-order logic is clearly a more powerful representation language than propositional logic, essentially because it allows to represent relations between objects. In propositional logic, each new statement requires a new proposition, while with a proper ontology first-order logic is much more compact and readable. However, predicate logic involves in general heavier manipulations, and it is preferable not to choose it when it is not needed. Chapters 4 and 5 will show problems which can be properly treated with propositional logic using PAS. Finally, relying on propositional logic, Sebastiani [Seb98] has underlined a misunderstanding in the use of logic for IR. Similarly, a clear representation of uncertainty *inside* (and not on top of) propositional logic, as is done with PAS, can bring a better understanding of the underlying assumptions when modeling the IR process.

3.4.4 The ”possible worlds” approaches

The ”possible worlds” approaches are probably the most studied one among the logical approaches to IR. We briefly discuss two of these approaches.

Retrieval by logical imaging is to our knowledge the only logical approach that faced the realm of the TREC experiments, apart from the PAS approach to retrieval in hypertext described in the next chapter. This technique exploits the statistical relationships between terms to transfer the weight of a term not occurring in a document, to the most similar term(s) occurring in it. Once the probabilities of the terms occurring in the documents are updated, retrieval is made as usual. Remark that no terms are added, neither to the document nor to the query representation. Logical imaging is a theoretical framework for describing probability transfers between terms, and appears to be very different in ideology from the PAS approach.

Nie and Brisebois applied a fuzzy modal logic to query expansion [Nie96]. The expansion process is described as a sequence of transitions between worlds, where a ”world” corresponds to a query representation, and transitions can be done by using thesaurus knowledge. One theoretical advantage on this technique over the PAS approach is that the maximum inference length can be fixed as a ”parameter”. In the PAS approach, the inference length cannot be defined inside the model.

3.4.5 Other logical approaches

There are many other logical approaches, and there is no room here for making a comparison between PAS and all of them. In this section, we selected approaches which appear, from our point of view, to be the more related with the PAS approach. If one would like to ”locate” PAS among logical approaches to IR, PAS could be placed somewhere between approaches which put emphasis on the representation of symbolic knowledge, and approaches which put emphasis on the assessment of uncertainty. It can be applied rather easily to real problems without losing the inferential power of logic.

3.5 Conclusion

This chapter presented the core of the logical model, and brought different justifications regarding its adequacy to IR modeling. The next two chapters will present applications of the model, which will be adapted or sometimes simplified to handle the case at hand.

Chapter 4

Retrieval in hypertext

This chapter addresses the problem of using hypertext links to help retrieval or organization of information. The chapter starts with a discussion on the approaches taken in IR to use hyperlink information, and on the new approaches which are now being developed on the Web (Section 4.1). The PAS model developed in the preceding chapter will be adapted to the present case (Section 4.2), and will be extended to compute a popularity measure or to find hubs and authorities given a well-represented search topic (Section 4.3). Moreover, special techniques for assessing the required probabilities must be developed (Section 4.4). Experiments on the small Web track (WT) and on the CACM collection will be presented (Section 4.5). A discussion on the potentiality of hyperlinks for retrieval will close this chapter (Section 4.6).

4.1 Approaches to retrieval in hypertext

With the widespread use of links in hypertext documents and Web pages, there is a growing interest in viewing hypertext links as additional sources of evidence about document content and relevance. This is illustrated by 1999 Trec'8 small Web track in which 17 groups participated: one of the specific questions addressed by the track was: "can links be used to enhance information retrieval?". However, the study of document relationships has not started with the Web: bibliographic references, for example, have been studied for a long time in IR and in infometrics. We will often use the generic term hyperlink to design all types of document relationships, because in our point of view there is no fundamental difference between these different types. Any document relationship can be described by a type, a source node and a destination node, eventually some text describing the "topic" of the link - the anchor. And for any collection of text containing document relationships stored in a computer, it is always possible to display the text with hypertext links corresponding to these document relationships, in order to allow browsing in the link structure. In our sense, the main difference between the different links will be the amount of information involved about relevance.

In the next two subsections, we report some of the main classical information retrieval approaches and the more recently developed Web based approaches to integrate the hypertext structure in retrieval.

4.1.1 Classical information retrieval approaches

Citations links can be found in various collections, most typically in collections of scientific articles where links between documents can be created on the basis of their bibliographic references, and in legal documentation where cases may refer to antecedents. Other document relationships can be generated starting from citation links : co-citation, bibliographic and Amsler links amongst others. We will not consider these links because they are implicitly contained in the original links. Other types of link are the Nearest Neighbors link (NN), established by computing a similarity between documents, and the relevance link, established between documents found relevant for a given request [Sav95]. Document relationships have been much studied as a source of evidence for document content or document relevance. There are several reasons for this:

- Citation indexing is independent of words and languages, and thus may partially remove the underlying ambiguity of natural language. For example, a relevant document which does not share (enough) index terms with the query can be retrieved because its is linked to some of the best ranked documents.
- Hyperlinks can be easily managed by computers because they usually follow a strict given pattern, at least in standardized collections. However this may not be always true on the Web.
- In some cases, links add new and unique capabilities. For example, before presenting a case as a precedent, the lawyer must make sure that a given decision has not been overruled, reversed, or limited in some way.

However, as nearly every other source of knowledge in IR, hypertext links represent a body of knowledge which has more or less reliability upon the circumstances. Linked documents may have vague or distant relationships, or only some parts of them may concern the same topics. And if a link can be valid for a request addressing the topics shared by the two documents, it can be misleading for requests which are not related to the shared topics. This leads to consider hyperlinks as probabilistic evidence that the documents are relevant together to the same information needs. For example, Martyn [Mar64, p.236] prefers to define bibliographic coupling (two documents sharing links to the same documents) as:

> ”an indication of the existence of the probability, value unknown, of relationships between two documents.”

The same way, Cleverdon questions the relevance of citation links [CMK66, Volume I, page 30]:

> ”It may be concluded that about half the references in an author’s paper are not included in connection with the main problem of the paper, a fact which may assist examination of the possibilities and limitations, of the bibliographic coupling and citation indexing.”

Roughly, approaches to the use of links in IR can be divided in two classes. A range of approaches assumes that links are evidence that two documents share similar contents. Links are then used to enhance document representations. Another range of approaches makes a further step: if a link is evidence that the two documents share similar content, then by extension, it is evidence that they should be relevant to the same request. Links can then be used to rerank documents after propagation of the document scores. We show now examples of the two approaches.

Links to modify document score and rank

In this type of approach, a link from a document d_1 to a document d_2 can be considered as evidence that the contents of d_2 might be (partly) similar to the contents of d_1. Consequently, if d_1 is relevant to a certain information need, then this can be considered as uncertain evidence that d_2 is also relevant to the same information need. In other words, d_1 being relevant increases the probability of d_2 being relevant. d_2 should then be placed at a better place in the ranking, reflecting this upgraded probability of being relevant.

The general scheme is to take an initial ranking and to rerank it as follows. The retrieval engine provides a ranking of documents based on their similarity with the query. It is assumed that the links have not been used to produce this ranking. In a second stage, the retrieval system takes account of the incoming and outcoming links related to the first m best-ranked documents (m can from 5 or 10 up to 200 documents, depending on the technique used). This way, the retrieval system must only consider the evidence brought by a small fraction of all available links.

There are different ways to implement this idea. One widely used is based on spreading activation [CT87]. Simply stated, the initial retrieval status value (RSV) of a document d, as handed out by the retrieval system, is updated by the weighted score of its m neighbors through a certain number of cycles. The neighbors can be citing but also cited documents. The RSV of d at cycle $i+1$ is computed by the following:

$$RSV(d^{i+1}) = RSV(d^i) + \sum_{j=1}^{m} \lambda_j RSV(d_j^i) \tag{4.1}$$

The parameter λ_j can be seen as the degree of certainty regarding the evidence provided by the incoming link from d_j to d. It can be a fixed value according to the link type, or may vary according to a measure of similarity between the documents and the query [FS95, Sav97].

The underlying assumption of a repeated propagation through a certain number of cycles is that documents may have direct but also indirect influences on each other through the links. Indeed, if the RSV of a document depends on the RSV of each of its neighbors, their RSV depend in turn on the RSV of their neighbors, and so on. In a way, this repeated propagation can be seen as an inference process, though with no guarantee that the inferences are always appropriate. However, the number of cycles is often limited to one; and more than one cycle is usually harmful to retrieval effectiveness [Sav97].

The spreading activation principle has at least two major advantages : (1) it can be easily understood and its application is straightforward, and (2) it can be added to the output of a retrieval system without any modification. However it has some weaknesses: (1) the values of parameters must be well chosen, and (2) the formula is mostly intuitive and cannot be justified by any model of IR.

Enhancing document content

The inference network model [TC91] considers document relationships as possible probabilistic dependencies. However, these dependencies cannot be directly encoded by links between documents in the inference network, because this would create the risk of having cycles, which are not allowed in inference nets. The document relationships are then used to enhance the content of documents. Take the set of terms T contained in a document d_1 that cites a document d_2. The method of Croft and Turtle

[CT93] consists in extending d_2's representation by (a) assigning a belief to the terms of T not initially contained in d_2, and (b) eventually increasing the belief of the terms of T already contained in d_2's representation. An improvement of 6% over the baseline is reported, but it must be said that the baseline can be considered very high for the CACM collection, on which experiments were done [CT93]. However, they conclude in a pessimistic way that the increase in retrieval effectiveness may not be worth the extra work in implementation, and the cost of nearly doubling the size of the inverted file (the file containing for each of the indexing terms, the documents that contain it).

Our modeling will both be influenced by the inference network and spreading activation modeling. As in the former, document relationships are considered by probabilistic dependencies. And as in the latter, these relationships are combined to an initial ranking of documents to compute a new (hopefully better) ranking.

4.1.2 Web-based approaches

Compared to standardized and delimited collections made of scientific articles, news stories, legal or medical documentation, the Web is quite disorienting in several respects. We review first some of the characteristics which make the Web so peculiar:

- **Size and evolution.** The Web is by far the largest collection of stored information, containing 800 million Web pages in date of February 1999 [LG99], and it is increasing at a rate of a few million Web pages a day. At the same time, it is estimated that 1% of the Web changes every week. At present time, we do not know for how much time will last its exponential progression [RRS98]. Search engines are for the moment "losing the race", as their relative coverage of the Web seems to diminish with time [LG99].

- **Indexing.** Web pages may vary from a few words to megabytes, such that classical indexing procedures (see Section 1.1.3) may produce unpredictable effect: it is reported in [BL98] that to a request on "Bill Clinton", a major search engine returned in first position a document containing only the words "Bill Clinton sucks" with an image of Bill Clinton (which is pretty natural regarding the way the vector space model works). Moreover, the assumption that indexing terms are good sources of evidence concerning the topics of a document is not as much verified as in more standard text collections, because often the choice of word is not innocent. People who design Web pages may have specific objectives in mind, which are reflected in their choice of words. For example it is reported in [otCP99] that the IBM company, which orientates itself towards a service company, avoids the use the word "computer" on its home page. And because of the commercial impact of having its site being often returned to the user (whether it is relevant or not to its information need), tactics to influence the ranking of search engines, called "spamming" or "the search engine persuasion problem" [Mar97] are widespread. In summary, one of the basic assumptions of text retrieval, that the concepts or topics of a document can be "told" from the terms it contains, is possibly less valid on the Web, at least in some cases.

- **Scope.** The kind of information (or sometimes disinformation) which can be found has virtually no limits: in a click, one can find commercial information, scientific papers, data repositories, propaganda or the Web pages of his friends. This seriously limits the potential use of techniques developed in IR for using domain knowledge, and aggravates the polysemy problem since words are often

used in every possible meaning. For example, a word such as ”parallel” is likely to have a specific meaning when found in collection of computer science articles, but on the Web it can take all imaginable significations.

- **Users and information needs.** Web users often do not know the principles to make an effective search and do not understand the underlying principles of search engines. They are often ”lazy” to formulate precisely their information need, and most queries are very short (one or two terms). Many users would expect to find relevant information from a request on ”cars”, while there are more than eleven million pages containing that term in Altavista’s index, and Yahoo! classifies the ”car” topic in more than 50 different subtopics. Moreover, there are very different types of information needs on the Web, which influence the way the request should be answered. For example, a user may be interested in commercial information on a product, may want a precise answer to a specific question (”What is the height of Mont Blanc?”), access to a set of resources for browsing, or find determinant information on a certain topic. Each of these types of information need requires a different type of answer.

- **Miscellaneous.** The Web has many specific characteristics which are not always found in more standard IR collection, such as mirror Web sites, text in multiple languages, little control of the HTML syntax by the Web pages creators, etc.

If the user has a pretty clear idea of the topic he is looking for, and if the topic is sufficiently represented on the Web, he may find what he is looking for in hierarchical classifications such as Yahoo!. More and more search engines come now with those classifications which were created and maintained up to now by humans. However the human cost of purely manual organization of information does not make it a viable option at long term, because the Web pages change continuously and new categories or ”cyber communities” appear everyday on the Web. And as reported in [Mar97], ”repositories are now themselves resorting to search engines to keep their database up-to-date”. Moreover, although the information can more easily be found, the user must still browse in a hierarchy to find it, and for many topics it is not so obvious where they should be classified in the taxonomy. Anyway, although these classifications are very helpful, there are a potentially infinite number of queries, which cannot obviously be handled by any classification. So, are we doomed to spend more and more time searching for information, as the Web gets bigger, more diverse and even more difficult to organize?

At present time it is not really possible to provide an answer to such a question, although there is an optimistic view that the improvement of resources will go faster than the growth of the Web. However there is a potentially very valuable source of knowledge for organizing and finding information: the few billion hypertext links which ”glues” the Internet. The Web would not be what it is without these links, which after all are the paths which lead to information. Browsing is for many users the usual way to find ”nearby” information, and as quoted in [Mar97]:

> ”The power of the Web resides in its capability of redirecting the information flow via hyperlinks, so it should appear natural that in order to evaluate the information content of a Web object, the Web structure has to be carefully analyzed.”

The IR community has acknowledged the potential but limited utility of the document relationships. The Web community seems to be very optimistic on the potential

usefulness of the hypertext structure to help organize and find information on the Web. This view is encouraged by recent experiments which seem to confirm that hyperlinks can be very valuable in locating or organizing information [Mar97, Kle98, BL98, CdBD99, Bh98]. And according to Chakrabati et al. [CdBD99, p. 550-551]:

> "Citations signify deliberate judgment by the page author. Although some fraction of citations are noisy, most citations are to semantically related material. Thus the relevance of a page is a reasonable indicator of the relevance of its neighbors, although the reliability of this rule falls off rapidly with increasing radius on average. Secondly, multiple citations from a single document are likely to cite semantically related documents as well."

We now review two of the most representative and popular approaches developed for handling hyperlinks in the context of the Web.

Estimating the popularity of a Web page

The PageRank algorithm, used in the Google search engine (www.google.com) considers that users have an absolute preference among Web pages: the more a Web page is visited, the more it is appreciated by the users. It is not possible to have access to the logs of the servers, but it is a reasonnable assumption that the preference of users is reflected in the hypertext structure: a link toward a Web page is often an indication that this page is acknowledged by someone as a good source of information. A simple way to implement this idea would be to count the number of times a Web page is cited. Microsoft's home page, surely one of the most visited page on the Web, is cited more than 23 million times in Altavista's index (probably much more in reality). However, each link should not be treated equally, since its impact also depends on the popularity of the parent node: a page cited only a few times but which is in Yahoo!'s index would certainly be quite popular. Thus the popularity of a page also depends on the popularity of the pages that cite it.

The idea behind PageRank is that a user who crawls the Web by selecting the hyperlinks at random is more likely to visit certain Web pages than others, simply because there are more possible ways by which the user can reach these pages. It is possible to model this as a Markov process, where the states of the system are each of the Web pages. To compute the matrix of transition probability, the assumption that the user will choose randomly one of the l_n outcoming links of the Web page: the transition probability to each of the destination pages is then $1/l_n$. The measure of popularity of a Web Page, its PageRank, is given by the stationary probabilities of this Markov process - the limit probability that the user will be on a certain page. The PageRanks are computed very simply by an iterative algorithm, which converges after a few steps.

This algorithm is criticized because it biases the access to information [LG99]. The "perverse" effect of PageRank is that it will push popular pages to get even more popular, and nearly unknown (unlinked) Web pages to stay unknown. As said in [Mar97], "visibility is likely to be a synonym of popularity, which is completely different than quality, and thus using it to gain higher score is a rather poor choice".

We will show how an alternative interpretation of popularity, more centered on the notion of relevance, can be implemented in the PAS approach.

Finding hubs and authorities

Up to now, we have always talked of documents being ”relevant” or ”non relevant” to an information need. However the Web is so large in scope that it is sometimes advisable to first cluster Web pages according to more general search topics, and to redirect the user into the cluster in which he will most likely answer his information need. This is the principle of the Yahoo! repository. There are many requests on the Web such as ”Mountain biking” or ”Human rights” for example, wherein a user is more interested in having good starting points for browsing in order to learn general information on these domains. Given a certain search topic, it is possible to distinguish two types of potentially ”relevant” pages: **authorities** and **hubs**. Authorities are pages which contain high quality and exhaustive information on the topic, and hubs are pages which contain links to the authorities, thus giving access to the information.

How can we find hubs and authorities? The idea of Kleinberg's HITS algorithm [Kle98] is to consider a root set of usually 200 documents, composed of the most likely relevant documents to a given topic. These 200 documents are found with a traditional search engine. This root set is expanded with all documents which point to or are pointed by these pages, to form the base set in which authorities and hubs will be found (the expansion can be done twice to have a larger base set). Then the connectivity of this base set is used to find the best hubs and authorities. The assumption is that a good authority is a page which has links from many good hubs, and a good hub is a page which has links towards many good authorities. The algorithm has some similarity with PageRank in that the quality of a page depends recursively on the quality of the neighbors.

This idea is implemented as follows. For each document d_p in the base set, a hub score h_p and a authority score a_p are computed. The initial scores are set to 1, but the final result is not sensitive to any non degenerate values of initial scores. Then the hub and authority scores are updated iteratively, by respectively the sum of authority scores of its child nodes and the sum of hub scores of its parent nodes. The updating equations are:

$$h_p = \sum_{d_p \to d_i} a_i$$

$$a_p = \sum_{d_i \to d_p} h_i$$

Kleinberg showed that the scores will converge if the scores are normalized after each step. The exact scores are not so important, since the user is presented with a ranked list of hubs and authorities: it is reported in [CDR+98] that after 5 steps the ordering of hubs and authority scores usually does not change anymore.

The algorithm has been improved since Kleinberg's initial proposal, by considering weighted sums of hubs and authority scores. three types of weight have been applied with positive effect, (a) a measure of appropriateness of the hyperlink measured by the number of words in its vicinity shared with the search topic [CDR+98], (b) a normalization by the number of links between the corresponding URLs of the source and destination nodes, to avoid mutually reinforcing relationships between two URLs [Bh98], (c) the initial similarity of the document to the search topic [Bh98].

We will see some of the weaknesses of this technique to estimate the hub and authority values of a Web page, and how a PAS modeling of this problem may provide a solution to it.

4.2 The hypertext retrieval model

4.2.1 Converting hyperlinks to knowledge

A link from document d_i to d_j is evidence that their contents is similar or related: from a hypertextual relationship, we wish to derive an informational relationship. It can then be interpreted as evidence that a user possessing the infons contained in d_i will possess the infons contained in d_j, and similarly that a user possessing the infons contained in d_j will possess the infons contained in d_i. In that sight it is natural to convert a hyperlink from d_i to d_j to:

$$\begin{aligned} D_i \wedge l_{ij} &\rightarrow D_j \\ D_j \wedge l_{ji} &\rightarrow D_i \end{aligned} \tag{4.2}$$

Previous research has demonstrated that following the links backward may provide a source of information of comparable value, and this is the reason why the link induces also a rule in the backward direction. The assumption l_{ij} can be understood as the uncertain condition under which the hyperlink from d_i to d_j implies an informational link, which holds with probability $p(l_{ij})$. This probability can be seen as an indicator of the quality of the link. If the condition holds, then we will consider the hyperlink as being "valid".

It is also possible to make the passage from a hypertextual to an informational relationship dependent on certain external conditions, such as certain concepts:

$$C_i \wedge C_j \rightarrow (D_k \wedge l_{kl} \rightarrow D_l) \tag{4.3}$$

$$C_i \wedge C_j \wedge D_k \wedge l_{kl} \rightarrow D_l \tag{4.4}$$

This kind of relationship is useful to introduce the context in the application of hyperlink knowledge. For example, suppose that d_k is a document about 'ski resorts in Europe', with a hyperlink towards d_l, a document about 'ski stations in Switzerland'. The hyperlink appears adequate if the user seeks documents about 'skiing in Switzerland', but not if he seeks documents about 'skiing in France'. If the concepts C_i and C_j denote respectively the concepts of 'ski' and 'Switzerland', the rule $C_i \wedge C_j \wedge D_k \wedge l_{kl} \rightarrow D_l$ will apply only for the query 'skiing in Switzerland'.

The probabilities $p(l_{ij})$ and $p(l_{ji})$ are assessed in a way that will be explained later. We now see the various ways in which this knowledge can be interpreted, depending on the way we want to use it.

4.2.2 Enhancing document content

If the knowledge induced by the links is included in the general PAS model where concepts are designated by the symbols C_i, their effect will be equivalent to enhancing document content. Suppose that there is a link from D_i to D_j ($D_i \wedge l_{ij} \rightarrow D_j$), and that D_j is about C_k ($D_j \wedge a_{jk} \rightarrow C_k$). By resolving the two rules, it follows that: $D_i \wedge l_{ij} \wedge a_{jk} \rightarrow C_k$. Then there are two possible cases. If there is no other (textual) evidence that D_i is about C_k, the support for C_k given D_i as evidence is: $sp(C_k, \xi \wedge D_i) = l_{ij} \wedge a_{jk}$. If there is a textual evidence that D_i is about C_k, represented by $D_i \wedge a_{ik} \rightarrow C_k$, then the support C_k becomes: $sp(C_k, \xi \wedge D_i) = a_{ik} \vee (l_{ij} \wedge a_{jk})$.

The effect of links is clearly to add new concepts in the description of the document, or to increase the weight of existing ones. This approach is quite similar to the way hyperlinks are taken into account in the inference network approach [CT93]. However in this approach the links are not interpreted in a transparent way. On the other hand, logic is not really indispensable if the goal is only to extend a document representation.

4.2.3 Modifying the rank

Take an information need represented by a proposition Q, and a retrieval system which outputs a ranking of documents to Q. If the links have not been taken into account in the computation, then it is possible to use them to post process this ranking, in order to have a hopefully better ranking of documents. In that sight, the initial ranking can be interpreted as uncertain evidence that the information need is about the document, or that the document is about the information need:

$$a_i \rightarrow (Q \rightarrow D_i) \tag{4.5}$$
$$b_i \rightarrow (D_i \rightarrow Q) \tag{4.6}$$

The probabilities $p(b_i)$ and $p(a_i)$ that the retrieval system "was right" in retrieving document d_i depend on the rank or score of the retrieved document. For example, the initial evidence on a document ranked first is stronger than on a document ranked 100th.

There are two ways of taking the knowledge induced by hypertext links into account, whether we use the $D \rightarrow Q$ or the $Q \rightarrow D$ approach. Since a hyperlink from d_1 to d_2 produces a rule from d_1 to d_2 and one from d_2 to d_1, then it can be seen easily that the two approaches lead to an equivalent form for the symbolic support of documents: to find the support $sp(Q, \xi \wedge D)$ from $sp(D, \xi \wedge Q)$, one must only change the a_i's by b_i's, and the l_{ij} by l_{ji}. However the numerical result can be different if the probabilities assigned to the replaced assumptions are different.

There is no conceptual difference in applying either the $D \rightarrow Q$ or the $Q \rightarrow D$ interpretation of relevance. But for the following we prefer to adopt a simpler and more natural formulation. In that formulation proposition D_i means "document d_i is relevant". The proposition D_i is then related to a clearly defined event, relevance, and the proposition Q is not needed[1]. The uncertain evidence provided by the retrieval system is modeled as some uncertain a priori knowledge on the relevance of the document: $a_i \rightarrow D_i$. The rules to represent hyperlink knowledge remain the same. The proposition Q is not needed anymore. To evaluate a document's relevance, we compute its symbolic and numerical support $sp(D, \xi)$ and $dsp(D, \xi)$. Next subsection shows an example.

Dealing with cycles

To illustrate how PAS deal with cycles in the hypertext network, take the following PAS: $\Sigma = \{a_1 \rightarrow D_1, a_2 \rightarrow D_2, a_3 \rightarrow D_3, D_1 \wedge l_{12} \rightarrow D_2, D_2 \wedge l_{23} \rightarrow D_3, D_3 \wedge l_{31} \rightarrow D_1\}$. Clearly, there is a cycle $d_1 - d_2 - d_3$. Does PAS deal with such a cycle, and avoids infinitely circular arguments? Take the case of D_1. The support of D_1 is: $sp(D_1, \xi) = a_1 \vee (a_3 \wedge l_{31}) \vee (a_2 \wedge l_{23} \wedge l_{31})$. One might ask why argument

[1] We could have used a new symbol such as R_i, but for the rest of this chapter, D_i will keep the same meaning, such that there should be no confusion for the reader.

$a_3 \wedge l_{31} \wedge l_{12} \wedge l_{23} \wedge l_{31}$, corresponding to argument $a_3 \wedge l_{31}$ with one cycle, would not be an argument for D_1? In fact, it is! But it is implicitly "contained" in argument $a_3 \wedge l_{31}$: $(a_3 \wedge l_3 1) \vee (a_3 \wedge l_{31} \wedge l_{12} \wedge l_{23} \wedge l_{31}) = a_3 \wedge l_{31}$, such that the symbolic support would not be modified by the addition of this "cyclic" argument.

More generally, if α is a minimal argument for hypothesis h (see Chapter 2) and $\alpha \wedge c$ is the same argument with one or more cycles, we have: $\alpha \vee (\alpha \wedge c) = \alpha$, such that the addition of $\alpha \wedge c$ will not modify the symbolic support. In a sense, the support can be said to contain implicitly all "cyclic" arguments.

4.3 Extensions to the model

This section proposes two extensions to the model, to compute a measure of popularity of documents, or to find hubs and authorities given a well represented topic in the hypertext collection. The models have not been experimented, but they illustrate that the PAS approach can serve for processing the hyperlinks in a variety of ways, and lead to a better understanding of the intuition behind a certain way of processing hyperlinks.

4.3.1 A model for computing a personalized "popularity" measure

The PageRank algorithm makes the hypothesis that the probability to be on a given Web page when crawling randomly on the Web is a measure of its popularity. We have a slightly different idea of "popularity" in mind: to our advice, the frequency at which a page is visited is not necessarily an indicator of its relevance, especially to someone who looks for a precise piece of information. The Microsoft front page, possibly the most upgraded by the PageRank algorithm, will not be of any interest for many users.

It is nevertheless interesting to study if the connectivity of a page contains some information on its a priori probability of being relevant to an information need. In that purpose, we can keep the model developed in the preceding section, and compute the symbolic support of each document. The rules inferred from following the hyperlinks backward should be kept, since they do tell us something about the relevance of documents (see Section 4.4). Remark that the PageRank algorithm does not consider links backward, on the assumption that the user does not have access (in general) to the incoming links of a Web page.

From this point of view, it seems appropriate to define P_i as "document d_i is popular", with the equivalent rules: $a_i \rightarrow P_i$ and $P_i \wedge l_{ij} \rightarrow P_j$. The support $sp(P_i, \xi)$ can then be interpreted as a symbolic measure of the popularity of the page d_i. The probability of each assumption are parameters the value of which can be modulated to fit at best the user's interest. For example, it is possible to evaluate separately the effect of the neighbors and indirect neighbors on the popularity of a page.

4.3.2 A model for computing hub and authority scores

In Kleinberg's algorithm, the authority/hub score of a document depends on the hub/authority scores of its neighbors. In a sense, the algorithm has an "objective" justification because it finds some intrinsic properties of a set of linked pages. However, we believe there are some weaknesses in this algorithm, which are not apparent because the way the algorithm works is not transparent, notwithstanding a certain mathematical cleanliness. A logical modeling of the problem can bring some insights on these weaknesses,

and on the way they can be solved. The most obvious weakness is that (1) although supposedly "objective", the final result depends on the base set (the chosen subset of the full graph), which in turn depends on the initial subset chosen, which is arbitrarily set to the 200 best ranked documents given by a certain search engine. This makes a lot of arbitrary choices for an "objective" measure. Since no comparative study has been published, we do not know if changing these degrees of freedom will affect greatly the final result. Also, (2) the initial ranking of documents is not used as prior evidence, while clearly an initially better ranked documents has more chances to be relevant [2]. And (3) although the weight of each link can be modulated, it is not clear if a simple addition of the contribution of each neighbor is appropriate.

In the modeling with the PAS framework, we assume that being a good hub or authority can be modeled by a binary variable. This can be supported by the fact that the binary relevance scale is privileged by the IR community, among other proposed relevance scales. For a document d_i, proposition H_i denotes "document d_i is a good hub" and A_i denotes "document d_i is a good authority".

We consider that there is initial evidence h_i that D_i is a good hub, and a_i that it is a good authority.

$$h_i \rightarrow H_i, a_i \rightarrow A_i \tag{4.7}$$

The probabilities $p(h_i)$ and $p(a_i)$ can be assessed with training data. The estimation can be made using a logistic regression, in a way similar to the technique shown in Section 4.4.

As in Kleinberg's algorithm, we make the assumption that if a document d_i is cited by a good hub d_j, then this is evidence that d_i is a good authority. We have then:

$$H_j \wedge f_{ji} \rightarrow A_i \tag{4.8}$$

where f_{ij} denotes the proposition "document d_i is a good authority if document d_j is a good hub". And similarly, if a good authority d_i is cited by a document d_j, then this is evidence that that d_j is a good hub:

$$A_i \wedge g_{ij} \rightarrow H_j \tag{4.9}$$

where f_{ij} denotes the proposition "document d_j is a good hub if document d_i is a good authority".

With this model, there is no need to determine a base set. It has not been implemented yet: although authorities can be loosely assimilated to documents judged relevant, one difficulty is that there is no test collection which makes the distinction between hubs and authorities. The few evaluating experiments done in this area do not respect the rigorous IR standard and prevent definite conclusions to be made.

4.4 Computing probabilities

Documents are ranked according to their degree of support, which depends on the probability given to assumptions. Correct estimates of these probabilities is then crucial. We show now a possible way to estimate these probabilities, with some results obtained from using the CACM and WT collections.

[2] The initial document score is used in [Bh98], but indirectly.

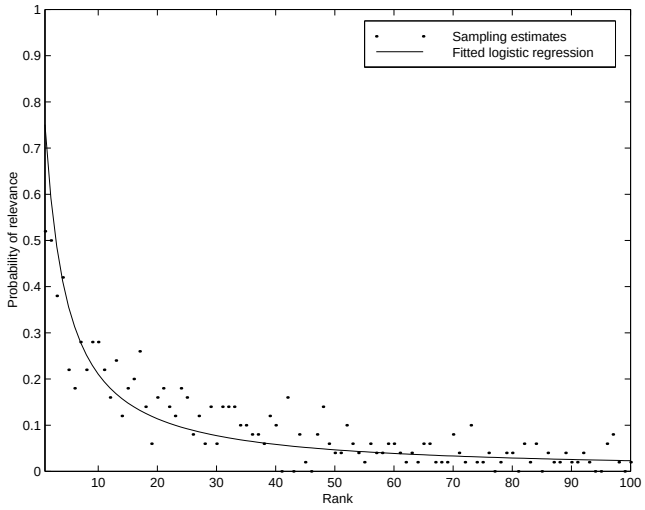

Figure 4.1: Probability of relevance vs Rank. CACM collection.

4.4.1 A priori support

To compute the probability $p(a_i)$ in $a_i \rightarrow D_i$, where D_i means "document d_i is relevant", we make use of some of the information handed out by the retrieval process, such as each of the document's rank r_i or score s_i. To estimate this probability, consider that if documents were not connected, a_i would be the only "argument" for d_i and clearly: $sp(D_i, \xi) = a_i$, and $dsp(D_i, \xi) = p(a_i)$. Documents would then be ranked according to $p(a_i)$.

It seems then natural to assess $p(a_i)$ on the basis of the evidence handed out by the retrieval system: $p(a_i)$ is estimated by $p(d_i = relevant|s_i, r_i)$. This probability can be assessed by fitting a logistic regression to a sample of training queries [CS00]. In our experiments with the CACM and WT collection, we used only rank as the explicating variable, because the probability of relevance is more stable with the rank than with the score. Figure 4.1 shows sampling estimates for the probability of relevance given the rank, and the fitted logistic regression. The score may provide additional information of interest, but this feature is highly correlated with the rank. A probability of 0 is assigned to a document not being retrieved.

4.4.2 Computing link assumption probabilities

In the uncertain rule $D_i \wedge l_{ij} \rightarrow D_j$, $p(l_{ij})$ represents the uncertainty associated with the knowledge that (the relevance of) D_j can be inferred from (the relevance of) D_i, and $p(l_{ij})$ represents the uncertainty associated with this event. From that viewpoint, a probability of 0.2 would mean that in 20% of the requests where d_i is relevant, d_j is also relevant. It is then natural to associate $p(l_{ij})$ with the conditional probability $p(d_j = relevant|d_i = relevant)$. This probability can be assessed on a sample set of queries with relevant documents, by computing the fraction of time that a document linked to a relevant document is in itself relevant. Obviously, the higher the probability of document relevance, the greater the link's information about relevance. The compu-

Estimation method	Incoming links	Outgoing links
Algorithm 1	0.145	0.106
Algorithm 2	0.066	0.090
Algorithm 3	0.062	0.051

Table 4.1: Probability estimation of hyperlinks- WT collection.

tation of this probability also gives an idea on what can (and cannot) be expected from links.

A possible technique for estimating this link probability is described below. Based on a set of queries along with their relevance assessments, we compute for each relevant document the fraction of linked documents that are themselves relevant, and then we compute the average of this fraction for all queries (Algorithm 1). An objection to this method is that some documents are linked to more than one relevant document, and thus will have a higher probability of being relevant. To avoid an overly biased estimate, we exclude these documents from the computation, and compute the probability in the same way as Algorithm 1 (Algorithm 2). Finally, the link probability might vary largely between queries, mostly because the number of relevant documents can also vary by one or even two orders of magnitude. In order to keep a few queries from dominating the computation, we take Algorithm 2 but compute the median instead of the mean (Algorithm 3).

From Table 4.1, one can find that depending on the algorithm used, the estimate may vary greatly. The experiments presented below make direct use of this probability, and work better for the smallest estimates found with Algorithm 3. This finding suggests that this value is a better estimate of the link's probability. It is lower than equivalent estimates found with the CACM collection, which can be explained by the fact that bibliographic references which have some intellectual justification, contain higher quality information than hyperlinks.

4.4.3 Semantic link probabilities

Fixed probabilities may not be very satisfying, because clearly some links are in general more appropriate than others, and links are more or less appropriate depending on the context (the context is the user's information need, represented by the query). We have adapted a technique developed by Frei and Stieger [FS95] which deals with the semantic of the links, to the computation of individual link probabilities. Chakrabati et al. [CDR+98] have also considered the similarity between the anchor of the hyperlink and the request when propagating hub and authority scores. We may regard here the similarity as a variable which affects the probability that the link will imply relevance.

The computation of this probability should take into account two factors: (1) the general appropriateness of the link, and (2) its adequacy to the present query. Let us denote $sim(d_1, d_2)$ as the cosine similarity between documents d_1 and d_2, computed from the intersection between their indexed representation. Also, $sim((d_1, d_2), q)$ denotes the cosine similarity between the link from d_1 to d_2, represented by their combined indexes, and the representation of the query q.

Intuitively, the general appropriateness of a link between two documents d_1 and d_2 should depend on the degree to which they share similar terms, which is measured by $sim(d_1, d_2)$. And the appropriateness of the link for a precise query should depend on the degree to which the query and the link between between the two documents share

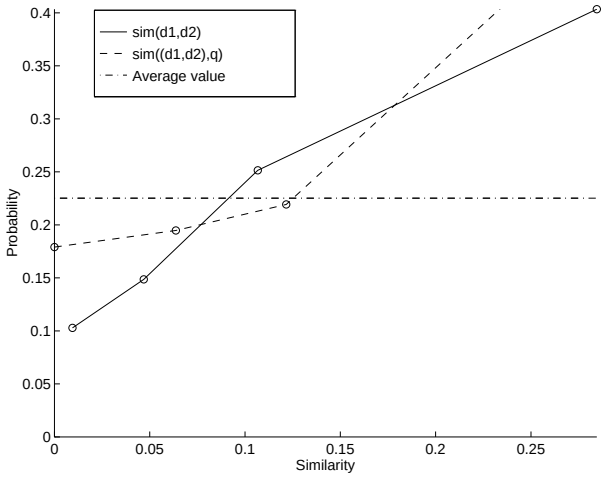

Figure 4.2: $P(d_j = relevant|d_i = relevant)$ vs $sim(d_1, d_2)$ and $sim((d_1, d_2), q)$. "Cited" links, CACM collection.

similar terms, which is measured by $sim((d_1, d_2), q)$.

In our experiments, links were represented by the ten most representative keywords of each of the documents. For each link from a relevant document to another document, the two cosine similarities are computed. As explained before, we consider only documents linked to one and only one relevant document. On Figure 4.2, the probability that d_j is relevant if d_i is relevant, $p(d_j = relevant|d_i = relevant)$, is computed for four sets of binned similarities. Clearly, the probability of a link assumption varies with the static and dynamic value of a link. It seems more appropriate to take as estimate of $p(a_i)$, a function $f(.)$ of $sim(d_1, d_2)$ or $sim((d_1, d_2), q)$, eventually of both. In future work, we will experiment with values of $p(a_i)$ parameterized by these two similarities.

4.5 Experiments

In this subsection, we present our experiments done on the CACM and WT collections.

4.5.1 Learning

The set of symbolic operations is as follows:

- Convert each bibliographic link from d_i to d_j to Eqn. (4.2).
- For each document D_i, compute the support $sp(D_i, \xi)$.
- Put the support of each document in disjoint form using Heidtmann's algorithm [Hei89]. Convert this disjoint form to a directly usable formula for computing the degree of support $dsp(D, \xi)$.

At this point, all the logical operations are done. When a query is processed, the only operation remaining is to compute the degree of support for each document

Type of link	SA	PAS	PAS vs baseline	PAS vs SA
Citing	0.266 (0.3)	0.267	+5.57%	+0.01%
Cited	0.261 (0.4)	0.273	+7.83%	+4.11%

Table 4.2: Experimental results on the CACM collection

$dsp(D, \xi)$. Values are assigned to the a priori assumption probabilities using the fitted logistic regression. If the probabilities of link assumptions are not static, then values must also be assigned depending on some similarity value between the link and the search topic.

4.5.2 Experiments on the CACM collection

In these experiments, links were considered separately in the forward direction (citing) and in the backward direction (cited). For each document, arguments of order three and less were computed, and the symbolic support was put in disjoint form with Heidtman's algorithm.

The basic retrieval process was done using a classical retrieval system based on the cosine similarity measure. The baseline retrieval effectiveness at 11 recall points is 25.27%, with the `TRECeval` software. Comparisons were made between PAS and baseline, and PAS and spreading activation (SA). The technique of spreading activation was described in Section 4.1.1. The results shown for spreading activation are the best ones obtained for a range of values of the parameter λ, which was fixed for all links of a certain type. Score was spread for only one cycle, since more than one was harmful to retrieval. The best λ value is shown between parenthesis. The results are shown in Table 4.2. In the PAS technique, arguments of length three or less (at most three links assumptions) were computed.

These results show that PAS can compete with an established technique such as spreading activation. Also, indirect neighbors can be considered without depreciating performance. For citing links, there is no average difference in retrieval effectiveness, while for cited links, the difference is nearly significant. However, for spreading activation there is no understanding of the parameters involved, while in the PAS framework, parameters are the probabilities of the links which have a clearer meaning both from a statistical and a logical viewpoint.

4.5.3 Experiments on the small Web Track

In a second set of experiments, the hyperlinks of the WT collection were used to improve retrieval. Ten different weighting schemes were used (okapi-npn, ...,bnn-bnn, see Table 4.3), to produce ten different rankings of documents. We do not describe these retrieval systems here. A simplified version of the PAS was used whereby a document's degree of support can be affected only by its direct neighbors. In that case we do not need to keep track of inferences, and can derive a simple formula which can be understood as a more refined way of spreading activation: instead of propagating a document's score, its probability of being relevant is propagated. This probability of relevance is estimated by the logistic regression technique. For the probabilities of the links, the lowest probability estimates found with Algorithm 3 are used (see Table 4.1).

In that case, the document's support is simply:

Model	Baseline	Best incoming	Best outcoming	Combined
okapi-npn	0.267	0.267 (0.00%)	0.267 (0.00%)	0.267 (0.00%)
lnu-ltc	0.234	0.239 (+2.18%)	0.240 (+2.65%)	0.239 (+2.10%)
atn-ntc	0.257	0.260 (+1.44%)	0.261 (+1.52%)	0.260 (+1.32%)
ntc-ntc	0.138	0.139 (+0.14%)	0.139 (+0.14%)	0.138 (0.00%)
ltc-ltc	0.136	0.138 (+0.88%)	0.138 (+0.88%)	0.138 (+0.88%)
lnc-ltc	0.107	0.108 (+0.65%)	0.109 (+1.49%)	0.107 (+1.31%)
lnc-lnc	0.072	0.074 (+2.35%)	0.073 (+1.38%)	0.073 (+1.52%)
anc-ltc	0.082	0.084 (+2.31%)	0.087 (+5.59%)	0.084 (+2.79%)
nnn-nnn	0.071	0.072 (+0.56%)	0.072 (+0.42%)	0.070 (-0.98%)
bnn-bnn	0.096	0.100 (+4.18%)	0.101 (+5.02%)	0.099 (+3.56%)

Table 4.3: Experimental results on the WT collection

$$sp(D,\xi) = a_i \bigvee_{link(D_k,d_i)} (a_k \wedge l_{ki}) \tag{4.10}$$

From the equality:

$$\begin{aligned} p(c_1 \vee ... \vee c_n) &= 1 - p(\neg(c_1 \vee ... \vee c_n)) && (4.11)\\ &= 1 - p(\neg c_1 \wedge ... \wedge \neg c_n) && (4.12)\\ &= 1 - p(\neg c_1) \cdot ... \cdot p(\neg c_n) && (4.13)\\ &= 1 - (1 - p(c_1)) \cdot ... \cdot (1 - p(c_n)) && (4.14) \end{aligned}$$

it follows that:

$$\begin{aligned} dsp(D_i,\xi) &= p(a_i \bigvee_{link(D_k,d_i)} (a_k \wedge l_{ki})) && (4.15)\\ &= 1 - (1 - p(a_i)) \cdot \prod_{link(d_k,d_i)} 1 - (p(a_k) \cdot p(l_{ki})) && (4.16) \end{aligned}$$

A first set of experiments using this formula for ranking documents by decreasing degree of support did not produce any increase of retrieval effectiveness, for the 10 weighting schemes used and different values of the probability of the link. This tends to show that simple and intuitive techniques, which have produced satisfactory results in other retrieval environments, do not seem to perform well on the Web. It is our opinion that hyperlinks seem to provide less information than do the bibliographic references or co-citation schemes used in our previous studies.

We have however obtained better results in some cases by considering only the best source of evidence for a document. That is, if a document receives evidence from three documents with ranks 3, 11 and 17, we only consider the evidence from document ranked 3rd. The reason for using only the best source of evidence is that when a document is already linked to one of the best-ranked documents, the other linked documents only have a marginal effect on its relevance probability. Table 4.3 show the retrieval effectiveness using only the best incoming links, the best outcoming links, and both. There are slight but generally not significant improvements over the baseline.

4.6 Conclusion

This chapter has presented an application of PAS to handle hypertext links for helping retrieval, eventually organization of information. The experiments have shown that logical models can be applied to a large collection for dealing with hyperlinks. It has also be shown that logic can be seen as tool which can be adapted to roughly any retrieval system, rather than being the ”cornerstone” of the retrieval system.

The use of links did not lead to very significant increase of retrieval effectiveness with the WT collection. Though it is hard to say what could at best be obtained from hyperlinks, it is doubtful that this information is of a very high quality. As an example, consider the estimated probability of 0.06 that a document linked to a relevant document is itself relevant: it can be interpreted as having roughly 15 misleading links out of 16. It seems that hyperlinks, although very abundant, are not always reliable information. The fact that a Web hyperlink contains less information than a bibliographic link can also be justified qualitatively: a user designing a Web page can add any hyperlinks, having different motivations in mind, while the author of a scientific paper has in general strong intellectual justifications for citing another paper.

Chapter 5

Retrieval with thesaurus

This chapter deals with the problem of implementing a retrieval system based on the PAS model developed in Chapter 3, with a focus on the integration of relationships coming from different thesauri. First, the thesaurus approach to IR will be presented, and the techniques or ideas used to create and take account of the term relationships will be described (Section 5.1). A new and original way of using term relationships will then be presented and justified (Section 5.2). Then different adaptations of the PAS logical model to take into account this information will be proposed (Section 5.3). Some experiments will be done on the implementation of a full PAS logical model of IR (Section 5.5). A discussion will end this chapter (Section 5.6).

5.1 Thesaurus and information retrieval

5.1.1 Manual thesaurus

A thesaurus is constituted by a set of keywords or concepts representative of a given domain of knowledge, and a set of relationships linking these different concepts, e.g. synonymy, hyponymy or hypernymy. The thesaurus can be general in scope, such as Wordnet and the Roget's which attempt to organize general human knowledge, or focus only a specific domain. The thesaurus can be weakly structured (e.g. the Roget's groups words into different categories such as 'dissimilarity' or 'religion') or can be organized into one or more hierarchies (Wordnet). With the increasing need of thesaurus in various domain of science, and with the generally very fast evolution of specialized scientific vocabulary, the construction of thesaurus is more and more done in a semi-automatic way.

Wordnet

In Wordnet, the thesaurus used in this thesis, keywords (words, expressions) are grouped into **synsets** (synonym sets). A keyword may belong to different synsets depending on its meaning. The synsets are organized into different hierarchies. The most important hierarchy is the hypernym/hyponym, or 'broader term'/'narrower term'. Another interesting hierarchy is the meronym/holonym, or 'is part of'/'has parts'. Figure 5.1 shows the two synsets for the word 'computer', its direct hypernyms and hyponyms. For each word, the relationships considered are its synonyms in different synsets, and its direct

```
2 senses of computer
1.  computer, data processor, electronic computer, in-
formation processing system -- (a machine for performing
calculations automatically)
2.  calculator, reckoner, figurer, estimator, computer
-- (an expert at calculation (or at operating calculating
machines))
```

Figure 5.1: Synsets of 'computer'

```
computer calculator synsn
computer chip meron
computer crt meron
computer diskette meron
computer estimator synsn
computer expert hypen
computer floppy meron
computer hardware meron
computer keyboard meron
computer machine hypen
computer mainframe meron
computer microchip meron
computer monitor meron
computer processor meron
```

Figure 5.2: Terms related to 'computer'

hypernyms, hyponyms, holonyms and meronyms. Figure 5.2 shows the related term of 'computer' extracted from Wordnet.

5.1.2 Statistical thesaurus

To build automatically a thesaurus, one must generally compute some kind of statistical similarity between the terms. Usually the similarity is computed on a large amount of text (e.g. the collection) related to the domain of interest [Sma93]. A similarity generally reflects the degree to which two terms appear together in the same context. The context can be a sentence, a window of words, a paragraph or the full document. It is also possible to consider only the set of documents relevant to a given request. To obtain more pertinent relationships (at the risk of missing some of them), we prefer using the sentence as context.

If the context is relatively long (e.g. the document), it is often better to consider a weight reflecting the importance of the term into the context. But with a sentence as context, binary weights are also appropriate. In that case we need only n_i, n_j and n_{ij}, as respectively the number of contexts in which term t_i occur, term t_j occur, and t_i and t_j co-occur.

There are different possible measures of similarity. Kim and Choi [KC99] compared different measures: the Jaccard ($\frac{n_{ij}}{n_i+n_j-n_{ij}}$), Dice ($\frac{2 \cdot n_{ij}}{n_i+n_j}$), cosine ($\frac{n_{ij}}{\sqrt{n_i \cdot n_j}}$) and two other measures they proposed. From their experiments, it seems that the Jaccard, Dice and cosine perform equally well. Our experiments will use the cosine similar-

```
iraq 0.083
civilian 0.081
force 0.075
iraqi 0.07
buildup 0.069
soviet 0.068
war 0.059
personnel 0.055
saddam 0.052
commander 0.051
```

Figure 5.3: Highest statistical similarities with 'military'

ity, which is largely used in information retrieval. Their query expansion experiments also show that retrieval effectiveness generally peaks when ten terms are added, and is stable or decreases slightly afterwards. We will also add only the ten best terms.

More sophisticated approaches measure a similarity between terms based on the pattern of words with which they co-occur. In an earlier work on the CACM collection [Pic99], we also considered second order co-occurrence, where the similarity between two terms depends on their pattern of co-occurrence with all the other terms [SP97]. This way, two never co-occurring terms may be related, if they co-occur in general with the same words. This type of measure is however computationally expensive even for a medium size collection.

Figure 5.3 shows the most similar terms to 'military', computed on the Wall Street Journal collection. The similarities are low in general because if two words may frequently co-occur in the same document, they more rarely co-occur in the same sentence. It is interesting to notice that some of the most similar terms to 'military' are related to the Gulf war held in 1991: this is not very surprising since the WSJ is composed of news article written between 1990 and 1992. These kind of contextual relationships can hardly be captured with a general manual thesaurus.

Remark that the highly similar term 'force' is also a hypernym of 'military' in Wordnet. Having the term in the two thesauri is a stronger indication of the appropriateness of the relationship between 'military' and 'force' for the present collection. As we will see, combining evidence from different thesauri results in better quality term relationships.

5.1.3 Using thesauri for IR

The use of thesaurus in IR is a very broad subject, and it is not our intention to make a detailed survey of the domain. Thesaurus are part of many retrieval systems, where they provide a useful help to the user in difficulty of formulating its information need. Here we are only concerned with the automatic use of thesaurus for retrieval. In general, the techniques which use thesaurus automatically for retrieval fall under the name of "query expansion": the terms in the initial query remain unchanged, and new terms are added to increase the possibilities of matching.

The association hypothesis [vR79] states that "if an index term is good at discriminating relevant from non-relevant documents, then any closely associated index term is also likely to be good at this". A "naive" interpretation of this hypothesis would be to simply add all the terms related to the terms in the query, such that a relevant

document containing only synonymous or related terms terms will be retrieved. This technique does not work in general, and thesauri must be used very carefully: increasing the chances of matching by adding related terms may help reducing the synonymy problem, but it nearly always worsens the polysemy problem.

Negative results using thesaurus for query expansion

Voorhees [Voo93, Voo94] made many experiments using the Wordnet thesaurus on the TREC corpus. Results were disappointing, even when the ”right” sense of a word was selected by hand. This led a part of the IR community to believe that the use of a general thesaurus is not a reliable technique to improve retrieval effectiveness. It is true indeed that there is a great challenge in attempting to apply the very general knowledge contained in Wordnet, which is not always appropriate to general purpose collections.

Smeaton and van Rijsbergen obtained disappointing results using statistical similarities to expand queries [SvR83]. They showed that selecting terms randomly does lead to worse retrieval results than choosing terms according to their similarity: this makes the quality of statistical relationships questionable. Later, Peat and Willet demonstrated that the cosine similarity emphasizes strong relationships between medium frequency terms, which tend to be poor discriminators of relevant and non-relevant documents [PW91].

Improvements on simple techniques

The recent improvements in the application of thesauri to query expansion lead to more optimism regarding their potential usefulness for IR. We survey the new ideas that are behind these improvements.

- **Considering the context.** A query term may have different meanings depending on the context, and of course it is not possible to tell the context from a term alone. But the context can be defined from the whole query. For example, in the two queries ”parallel skiing” and ”parallel computers”, the meaning of the term parallel can be told from the other term. Qiu and Frei developed a query expansion technique with a statistical thesaurus where terms are selected according to their average similarity with all the query terms [QF93]. This way, an inappropriate term which would be similar to only one term of the query would have a low average similarity and would not be retained. The idea of taking account of all the query terms for selecting the expansion terms seems well adopted now [XC96, RTT99, KC99].

- **Combining different thesauri.** Used alone, manual thesauri are not very helpful on general purpose collections. However Mandala et al. [RTT99] showed that when Wordnet is used in conjunction with other statistical thesauri, it can be a much more useful source of evidence. On TREC’7 data, they obtained 58.2% improvement using two statistical thesaurus, and 98.9% with Wordnet and these thesaurus. Wordnet alone led to only 10.2% improvement. However it must be said that the baseline was relatively low (0.1175 of mean average precision), and that the use of the thesauri only raised retrieval effectiveness to the level of standard retrieval systems that made the TREC competition.

- **Learning.** In the two preceding works, nice retrieval results were obtained using the co-occurrence similarity value without any modification. But there is no

reason that the optimal propagation value should be equal to the similarity. It is reasonable to believe that the obtained results would be even better if a more careful use of the similarities was made. Using a logical approach for expansion on the CACM with Wordnet, Nie used a learning technique on past relevance judgements to find adequate propagation values depending on the type of relationships (synonyms, hypernyms..) [Nie96]. Interestingly, he showed that if there is enough learning data, it is possible to obtain a propagation value for a precise relationship, as already told in Chapter 3. However, we do not know of any experiments for learning the propagation value from the statistical similarity.

5.1.4 Discussion

From what has been seen, it sounds clear what features should be those of a good query expansion technique: (1) considering all query terms to choose expansion terms , (2) combining different thesauri, and (3) learning the appropriate value for a relationship by using past relevance judgements. We will see how these ideas can be adapted to the PAS approach. But before, another point must be developed, concerning the problem of determining which query terms are appropriate for representing the information need.

5.2 Determining content-bearing query terms

The end of the previous section highlighted some of the features that should be integrated in a query expansion technique based on thesaurus knowledge. This section proposes a different way to take account of statistical similarities: they are not used here to add relevant descriptors of the information need to the query, but to determine the subset of query terms which are good descriptors of the user's information need, in order to improve the weighting of query terms. With the abstract features used to weight terms (query frequency qf and idf), this issue cannot be specifically addressed. Since to our knowledge, this topic has never been addressed in the information retrieval literature, it must be first justified and verified experimentally. We start by explaining and justifying what we name the "Cluster Hypothesis for query terms". Then we present and discuss the methodology adopted to verify this hypothesis.

5.2.1 The Cluster Hypothesis for query terms

Suppose that query terms can be divided roughly into those that are useful for retrieval and those that are harmful, which we will call respectively relevant or "content" terms, and non-relevant or "noisy" terms. In this section, we wish to test the hypothesis that two relevant terms tend to be statistically more similar to each other than would be two noisy terms, or a noisy and a content term. Intuitively, the terms which concern the topic of the query should in general concern similar topic areas. Consequently, they should be found in similar contexts in the corpus. A similarity measures the degree to which two terms can be found in the same context, and should be higher for two content terms. It could then be possible to use similarities between query terms to adjust the weights of the terms.

We propose to name the proposed hypothesis as the "Cluster Hypothesis for query terms", due to its correspondence with the famous Cluster Hypothesis of information retrieval which assumes that relevant documents "are more like one another than they are like non-relevant documents" [vRSJ73, p.252]. In a similar way, our hypothesis

Removed term	Ret. eff.
figure	0.2837
producer	0.2588
oil	0.17
total	0.2541
change	0.2801
specific	0.2464
natural	0.1451
gas	0.1547
provide	0.263
data	0.27
proven	0.1239
reserve	0.1315

Table 5.1: Retrieval effectiveness obtained after removal of each of the query terms (TREC topic #90). Retrieval effectiveness using all query terms is 0.2307. Useful terms are in bold .

assumes that relevant terms are more similar to each other than to noisy terms. This hypothesis can be summarized in three points:

- query terms relevant to the information need are in general more likely to concern similar topics;
- terms which concern similar topics should be found in similar contexts of the corpus (documents, sentences, neighboring words...);
- terms found in similar contexts have a high similarity value. Consequently, relevant terms tend to be similar to each other.

The following request (TREC topic #90) illustrates the hypothesis that similarities between query terms can serve as evidence for their relevance relatively to the user's information need:

> Document will provide totals or specific data on changes to the proven reserve figures for any oil or natural gas producer.

With the basic retrieval model which will be described in the next section, the initial average precision found for this topic on the WSJ collection was 0.2307. The average precision found after removal of each of the query terms is shown in Table 5.1. Clearly, a lower retrieval effectiveness without a term means that this term is a useful descriptor of the information need, and a higher one means that it is harmful.

From Table 5.1, it appears that the useful terms are 'oil', 'natural', 'gas', 'proven' and 'reserve'. It sounds clear that most of these terms are good descriptors of the information need[1]. The terms which degrade precision seem not to address specifically the user's information need. For example, 'figure', 'data' or 'change' can be found in

[1] Remark that we do not consider here phrases such as 'natural gas', but the hypothesis could be extended to phrases.

a wide range of topics. If the retrieval system could "know" which terms are useful descriptors, it could adjust their weights in consequence. The abstract features used to weight terms (query frequency and idf) are not sufficient for that purpose. A term appearing with the same frequency in two queries can be useful in one and harmful in the other, while its weight will be the same. In this query, the weight of 'figure' (0.216) is approximately the same as 'gas' (0.213), although the former appears to be the most harmful term and the latter is one the most useful.

A cosine similarity was computed between all query terms based on the WSJ collection, using each sentence as context. The highest similarities were: gas-natural (0.418), gas-oil (0.158), producer-oil (0.038), natural-reserve (0.037), oil-reserve (0.037), producer-oil (0.038) and proven-reserve (0.03). All the other similarities were lower than 0.02. It seems indeed that these terms concern similar topics, and this is reflected in a higher similarity. This cluster of similarities could be taken into account to determine the useful query terms.

5.2.2 Determining relevant terms

To verify the Cluster Hypothesis for query terms, we must have some way to determine the usefulness of a query term. Until now, we have assumed a binary classification between useful and harmful terms, which needs to be justified. It is obviously true that the use of a term will either increase or decrease retrieval effectiveness. However, some terms affect slightly retrieval effectiveness while others have a larger influence. Moreover, it is possible that by using a different retrieval system, weighting function, or even a different evaluation measure of retrieval effectiveness, the classification of the terms could be slightly different.

In a previous paper which addressed the statement and verification of this hypothesis on the CACM collection [Pic99], we used the χ^2 test of independence between the occurrence of the term and the relevance of the document to determine if the term is relevant or noisy. For each query term, a χ^2 value is computed based on the number of times the term (occurs/does not occur in (relevant/not relevant) documents. A threshold is associated to a certain confidence level: in our case, the threshold is 3.86, for a confidence level of 95%. If $\chi^2 > 3.86$, then the hypothesis of independence between occurrence of the term and relevance of the document is rejected at the 95% confidence level, and the term is considered relevant. Otherwise, it is considered noisy. This test has the advantage of being independent from any retrieval approach. However the unsatisfactory point with the χ^2 test is that it is not guaranteed that a term judged useful has indeed a positive effect on retrieval effectiveness. Since improving retrieval effectiveness is the ultimate goal, testing the usefulness of a query term by comparing the average precision with and without this term appears to be more appropriate. We will assume that in this "precision test", a term is relevant to the information need if the precision obtained without using that term is lower than with the original query, whatever the difference is. The weakness of this approach is that all terms are considered equally, whatever their effect their removal has on the query.

Table 5.2 makes a comparison between the retrieval effectiveness test and the χ^2 test of independence. In general, the retrieval effectiveness test is less "optimistic" than the χ^2 test, with 50.51% (744 on 1473) of terms classified relevant vs 61.37%. It is remarkable that one half of query terms should rather not be used for retrieval. The two tests make different classifications for 230+70=300 terms (20.3%). It seems that the χ^2 test is more flexible towards unfrequent terms: the average probability of occurrence $p(occ)$ of the 1473 terms is 0.055, while the 230 terms classified not relevant with the

	Lower precision	Higher precision	Total
$\chi^2 > 3.86$	674 (-0.062)	230 (+0.022)	804
$\chi^2 < 3.86$	70 (-0.008)	499 (+0.022)	569
Total	744	729	1473

Table 5.2: Comparison of χ^2 test and precision test. The average variation of precision is in parenthesis.

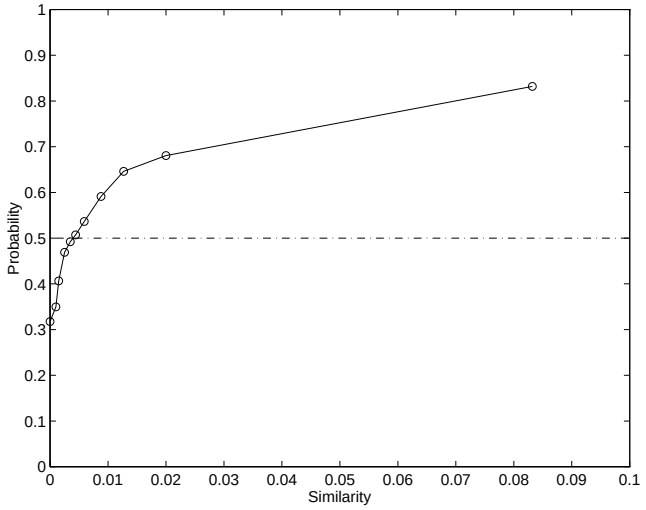

Figure 5.4: Probability that a query term is relevant vs Similarity with a relevant term. Dashed line (-.-): estimated a priori probability that term is relevant (=50.51 %).

χ^2 test but relevant with the retrieval effectiveness test have a probability of occurrence of 0.045. Moreover, the 70 terms classified not relevant with the χ^2 test but relevant with the precision test are very frequent: they have a probability of occurrence of 0.138.

It is interesting to compare the average variation of precision for each of the four possible classifications. The terms classified relevant with both tests seem to be very useful for precision (-0.066 without them). On the other hand, the terms not relevant with the precision test are equally harmful whether they are considered relevant or not with the χ^2 test. This suggests that the χ^2 does not discriminate between terms not relevant with the precision test. However the 70 terms rejected by this test but not by the precision test affect marginally precision (-0.008 without them).

5.2.3 Verification of the hypothesis

Figure 5.4 shows the estimated probability that a term is relevant given its similarity with a relevant term. A term is considered relevant according to the precision test. On the 185 queries of the WSJ collection, there were 2890 term pairs with at least one relevant term, of which 589 had a null similarity. The other 2301 pairs were divided

in ten bins of 230 pairs (except one with 231) by decreasing similarity. On the figure, each 'o' corresponds to one bin. The x value of a bin is the average similarity, and the y value is the frequency estimate of the probability that the other term of the pair is relevant.

Unequivocally, the probability that a term is relevant augments in a monotonic way with its similarity with a relevant terms. For a similarity of approximately 0.005, the probability gets over the a priori probability of 50.51%. It is interesting to notice that a very small or null probability could be used as evidence against the relevance of a term.

5.3 Proposed models

In this section are presented different adaptations of the general PAS model, to integrate term relationships in a computationally feasible way. We start by presenting the "basic" model which serves as a baseline. Then this PAS is extended to integrate thesaurus knowledge. Finally, a PAS is developed to take account of the similarities between the query terms in the computation of their weights.

5.3.1 Basic model

In the so called basic model, the knowledge base ξ_B does not include thesaurus knowledge. This PAS is similar to standard term matching approaches. It is used to find proper probability values for the relationships between concepts and information needs, and concepts and documents. It serves also as a benchmark for the PAS which will integrate thesaurus knowledge. In order to design this model, an interpretation of relevance ($D \rightarrow Q$ or $Q \rightarrow D$) must be chosen. In the $D \rightarrow Q$ interpretation of relevance, the starting point of inferences is the document D, while in the $Q \rightarrow D$ interpretation, the starting point is the information need Q. In some way, the two interpretations can be made symbolically equivalent if the order of rules is reversed. However, the concept relationships are used basically for query modification and expansion, in order to have a better description of the information need. Since the reasoning process is centered on the information need, we think it is more natural to take Q as "evidence". The $Q \rightarrow D$ interpretation will then be used, and the evaluation of a document's relevance is made by computing the symbolic support $sp(D, \xi_B \wedge Q)$ and the numerical degree of support $dsp(D, \xi_B \wedge Q)$.

The set of variables of interest is $P = \{D, Q, C_1, ..., C_p\}$. Since only one document and one information need are considered at once, it is not necessary to differentiate them withindices. In the basic model, only two types of rules are admitted: if a concept C_i is (probably) a descriptor of the information need, we have: $Q \wedge d_i \rightarrow C_i$. The same way, if it is a descriptor of the document, we have: $C_i \wedge b_i \rightarrow D$.

Without loss of generality, assume that concepts C_1 to C_m are the only descriptors common to the information need and the document. Then for i=1,..., m, $(d_i \wedge b_i)$ is an argument for hypothesis D. Clearly, the symbolic support of D is:

$$sp(D, \xi_B \wedge Q) = \bigvee_{i=1}^{m} (d_i \wedge b_i) \tag{5.1}$$

And following Equation 4.11 ,the degree of support is:

$$dsp(D, \xi_B \wedge Q) = 1 - \prod_{i=1}^{m}(1 - p(d_i) \cdot p(b_i)) \tag{5.2}$$

For small values of the degree of support, this ranking formula produces results very similar to the standard internal product:

$$score(D, Q) = \sum_{i=1}^{m} p(d_i) \cdot p(b_i) \tag{5.3}$$

The experiments will compare the mean average precision found with the two matching formulas. It seems that there were relatively few differences in the ranking of documents with the two matching functions, although no attempt was made to estimate a distance between the two corresponding rankings of documents.

5.3.2 Considering thesaurus information

To the basic knowledge base ξ_B just presented, we wish to add some relationships between concepts coming from a thesaurus, leading to a knowledge base ξ_T. For our experiments, this knowledge is always interpreted as rules of the following type: $C_i \wedge r_{ij} \rightarrow C_j$. If no restriction is put on the size of arguments, the kind of arguments we obtain with these rules is of the general form:$(d_h \wedge b_k \wedge r_{hi} \wedge ... \wedge r_{jk})$. For example, suppose that C_i is a descriptor of the information need ($Q \wedge d_i \rightarrow C_i$), that C_j has a link from C_i and C_k has a link from C_j ($C_i \wedge r_{ij} \rightarrow C_j$, $C_j \wedge r_{jk} \rightarrow C_k$), and finally that C_k is a descriptor of D ($C_k \wedge b_k \rightarrow D$). Then $(d_i \wedge b_k \wedge r_{ij} \wedge r_{jk})$ would be an argument for D. This argument could correspond to a document containing the synonym of the synonym of a query concept.

We will apply the following restriction: only arguments of size three or less are admitted. This avoids to include too long inferences in the computations. The eliminated arguments would anyway have a very small probability of being true, given the order of values we assign to the probabilities $p(b_i)$ and $p(r_{ij})$ (see next section). Moreover, the more an argument contains assumptions, the less it matches intuition. The synonym of a synonym of a query concept is unlikely to be relevant to the information need.

Assume without loss of generality that the concepts common to the information need and the document are denoted by C_1 to C_m. Of course, either the document and the information need may contain other concepts. The possibilities of matching can be extended by rules from any concepts representing the information need to any other concepts C_k of D. That is, we may have a sequence of rules of this type: $Q \wedge d_i \rightarrow C_i$, $C_i \wedge r_{ij} \rightarrow C_j$, $C_j \wedge b_j \rightarrow D$. Such a sequence, which leads to arguments of the type: $(d_i \wedge r_{ij} \wedge b_j)$, is denoted by $expansion$ in the following equation:

$$sp(D, Q \wedge \xi_T) = \bigvee_{i=1}^{m}(d_i \wedge b_i) \bigvee_{expansion} (d_i \wedge r_{ij} \wedge b_j) \tag{5.4}$$

The concept relationships may have two effects: (1) they increase the possibilities of matching between the document and the information need with an originally not shared concept, (2) and they (sometimes) enhance the weight of an original query concept. The support of a concept C_i representing the information need is initially: $sp(C_i, \xi_B \wedge Q) = d_i$. But if C_i is linked to another query concept C_j, it becomes: $sp(C_i, \xi_T \wedge Q) = d_i \vee (d_j \wedge r_{ij})$.

Using an inverted file for indexing and concept relationships, computing the symbolic support of each document is approximately linear with the number of query concepts. But to obtain an exact numerical evaluation of this support, it is necessary to put it in disjoint form. If Heidtmann's algorithm was used (as in Chapter 4), significant computational costs can be involved if many documents have a high number of arguments, because the algorithm is $O(n^2)$ with the number n of arguments. It would certainly be possible to design a faster algorithm optimized for this problem, but this is beyond the subject of this thesis. However, intuition suggests that very little improvements, if any, would be obtained by an exact computation of the symbolic support, and efficient approximations can be made.

The support $sp(D, Q \wedge \xi_T)$ is of the form $\alpha_1 \vee \alpha_n$, where the $alpha_i$'s are arguments. The approximation made will be to consider arguments α_1 to α_n as if they had no assumption in common: whatever i, j, $\alpha_i \wedge \alpha_j = F$. In that case , following Equation 4.11, the degree of support is:

$$dsp(D, Q \wedge \xi_T) = 1 - \prod_{i=1}^{n}(1 - p(\alpha_i)) \tag{5.5}$$

With this approximation, the degree of support $dsp(D, Q \wedge \xi_T)$ is computed with the following:

$$dsp(D, Q \wedge \xi_T) \simeq 1 - \prod_{i=1}^{m}(1 - p(d_i) \cdot p(b_i)) \cdot \prod_{expansion}(1 - p(d_i) \cdot p(r_{ij}) \cdot p(b_j)) \tag{5.6}$$

This kind of approximation may overestimate slightly the symbolic support, but it can be shown that the order of the bias is of $p(r_{ij})/100$ for the individual contribution of an argument, with typical values for $p(d_i)$ and $p(b_i)$.

5.3.3 A PAS for improving query term weights

In Section 5.2, we have seen that term similarities could be used to adjust query term weights, but it is not obvious how they can be used: the relevance of each query term to the information need depends on its similarity with the other query terms and on the relevance of these other query terms, which is unknown. As we will see, it seems that PAS are well adapted to deal with this type of evidence. This application of PAS departs from the logical model, in a way similar to the approach taken in Chapter 4 where the ranks of documents were adjusted after taking into account the document relationships. The variables of interest in the PAS refer to the query terms, and similarities are interpreted as uncertain rules linking the relevance to the information need of the concerned terms. Once a PAS is designed for a precise query, it is used to compute the degree of support that each query term is relevant to the information need. This degree of support is then used to increase or possibly decrease the weight of the query term, in order to obtain better retrieval effectiveness.

The needed variables are $\{T_1, .., T_n\}$, propositions associated with the terms t_1 to t_n contained in the query. The meaning of T_i is "term t_i is relevant to the information need". Each term in the query has an a priori probability of being relevant: it is represented by $a_i \rightarrow T_i$, with $p(a_i)$ as the a priori probability that term t_i is relevant[2]. We will see in the next section the frequency of a query term influences its probability of

[2]The assumptions a_i and l_{ij} used here must not be confounded with those of Chapters 3 and 4

relevance, and can be used to have a better estimate of $p(a_i)$. For the pairs of terms which have a significantly high similarity, the following rules will be integrated in the PAS:

$$T_i \wedge l_{ij} \rightarrow T_j, T_j \wedge l_{ji} \rightarrow T_i \tag{5.7}$$

where the main factor influencing the probabilities $p(l_{ij})$ and $p(l_{ji})$ is obviously the similarity $sim(t_i, t_j)$.

We have seen that for a term t_j, a low similarity (below 0.005) with a relevant term t_i is evidence that t_j is not relevant to the information need (see Figure 5.4). This can be modeled by the following rule:

$$T_i \wedge l_{ij} \rightarrow \neg T_j \tag{5.8}$$

However we will not consider negative evidence in our experiments.

Arguments of length more than three (which contain more than two link assumptions) are not included in the computation of the support. The support of hypothesis T_i will then be composed of three types of arguments:

- a_i: this is the initial evidence brought by the number of occurrences of t_i;
- $a_j \wedge l_{ji}$: this is the evidence brought by a high similarity value with query term t_j. However this argument only considers the a priori evidence that t_j is relevant, while other neighbors of t_j may affect its relevance, and consequently t_i's relevance;
- $a_k \wedge l_{kj} \wedge l_{ji}$: this type of argument allows to take account of the indirect evidence from t_j's neighbor t_k in the computation of T_i's degree of support.

Once the degree of support of each query term is computed, how can it be used to modify query term weights? In this thesis we will admit that the weight of the term is modified according to the heuristic formula 5.9. A more formal treatment of this problem is possible: a probability can be derived from the degrees of support and plausibility by using a pignistic transformation [SK93]. And the probability that a term is relevant to the information need affects its probability of occurrence in relevant documents, which in turn affect its weight according to the Robertson and Sparck-Jones weighting formula [RSJ76]. However we think that before attempting to try more sophisticated approaches, it is preferable to verify if encouraging results with a more primitive technique support further investigations.

The difference $\delta_i = (dsp(T_i, \xi) - p(a_i))$ is a measure of the degree to which the similarities with the other query terms support the fact that t_i is relevant to the information need. The value δ_i is equal to zero only when t_i is not at all supported by the other terms, in which case $sp(T_i, \xi) = a_i$ and $dsp(T_i, \xi) = p(a_i)$. We assume that the variation of the weight of term t_i depends linearly on its initial weight $old(t_i)$ and on $delta_i$. The new weight $new(t_i)$ of term t_i is then computed by:

$$new(t_i) = old(t_i) \cdot (1 + \beta \cdot \delta_i)) \tag{5.9}$$

where the parameter β is to be fixed in order to optimize a retrieval effectiveness criteria. The higher β, the more term similarities have an effect on query term weights.

Once the new weight of each query term is computed, the retrieval system in which this PAS is implanted can retrieve documents as usual. In the experiments, we will see how to determine the probabilities of assumptions and the value for the β parameter.

5.4 Experiments

5.4.1 Basic model

We need two types of probability estimates for the basic model.

- The probability $p(d_i)$ of inferring a concept C_i from the query (in $Q \wedge d_i \rightarrow C_i$);
- The probability $p(b_i)$ of inferring a document D from a concept C_i (in $C_i \wedge b_i \rightarrow D$).

For the estimation of these probabilities, we follow the technique taken with the inference network model [TC91]. That is, we assume that the probability $p(b_i)$ depends in the following way on the tf and idf values of the corresponding term :

$$p(b_i) = \alpha + (1 - \alpha) \cdot tf \cdot idf \tag{5.10}$$

where α is a parameter between 0 and 1 to optimize according to the retrieval effectiveness criteria. For the estimate of $p(d_i)$, we will take simply:

$$p(d_i) = \frac{qf}{qf_{max}} \tag{5.11}$$

where qf is the frequency of the term, and qf_{max} is the maximum frequency of a term in the query. Other estimates of $p(d_i)$ such as: $\alpha + (1-\alpha) \cdot \frac{qf}{qf_{max}}$, were attempted, but any value of α higher than 0 deteriorated retrieval effectiveness.

With these values of probabilities, the matching formula $\sum p(d_i) \cdot p(b_i)$ is equivalent to the inference network approach, apart from a normalizing constant.

Table 5.3 gives the mean average precision on the Wall Street Journal (WSJ) collection (185 requests), for various values of α with the PAS and the internal product matching formulas (IN). The range [0.05,0.2] was more densely explored since the most appropriate values of α seem to lie in this area.

The results are very similar for the range [0.1,0.2]. The value 0.175 appears to be a good estimate for α. The following formula gives the estimate of $p(b_i)$ that we will use for the rest of the experiments. This estimates will also be used for the internal product.

$$p(b_i) = 0.175 + 0.825 \cdot tf \cdot idf \tag{5.12}$$

It must be said that the relatively high baseline of 0.1995 taken for the experiments with the PAS (0.1971 for the internal product) will make it more difficult to achieve better retrieval effectiveness using thesaurus knowledge.

5.4.2 Use of Wordnet

It is generally recognized that the use of Wordnet thesaurus alone will not improve retrieval effectiveness on TREC data [Voo93, Voo94]. We wished to test whether each of the most important relationships between concepts (synonymy, meronymy, hypernymy and holonymy) could separately improve retrieval. Ony one type of relationship is used at once. The concept relationships were used according to the model developed in Section 5.2, with different values of $p(r_{ij})$ ranging from 0.1 to 1. Only the 10 new concepts which had the highest degree of support were added to the original query. Results are shown in Table 5.4.

α	PAS	IP
0	0.1489	0.1473
0.05	0.1819	0.1791
0.1	0.196	0.1954
0.125	0.1989	0.1995
0.15	0.1987	0.2004
0.175	0.1995	0.1971
0.2	0.1936	0.1961
0.3	0.1791	0.1845
0.4	0.1679	0.1757
0.5	0.1575	0.1645
0.6	0.1493	0.1554
0.7	0.1435	0.1503
0.8	0.1393	0.1474

Table 5.3: Mean average precision found for various values of parameter α. PAS basic model and internal product (IP) , WSJ collection.

For every type of relationship, the mean average precision degrades as the value of $p(r_{ij})$ increases. This confirms that the Wordnet thesaurus taken alone does not appear to be suitable in general for this collection, whether a logical or a vector-space approach is taken. However, this does not mean that this thesaurus is unusable, but more refined probability estimates must be given to concept relationships: the probabilities should not be uniform for all relationships of a certain type.

5.4.3 Statistical relationships

In this experiment, we wish to adapt Qiu and Frei's expansion technique to the PAS framework. We will use here an adaptation of their technique, with a parameter α added to weight the importance of the expansion terms compared to the original query term:

$$w(t_j) = \sum_{t_i \in query} \alpha \cdot q_i \cdot sim(t_j, t_i) \tag{5.13}$$

The candidate terms having the highest weights are kept for expansion (ten in our experiments). Remark that the parameter α does not modify the terms that are added to the query, but their relative importance.

In the adaptation to the PAS framework, a high similarity is converted to a rule $C_i \wedge r_{ij} \rightarrow C_j$. The conversion uses the same parameter α weighting the importance of the similarity value:

$$p(r_{ij}) = \alpha \cdot sim(t_i, t_j) \tag{5.14}$$

The ten candidate concepts which the highest degree of support $dsp(C_i, \xi \wedge Q)$ are added to the query. On the 185 requests, the terms that were added the most frequently with the PAS technique are: economic (12 times), bank (10), iraq (10), market (9), patient (9), price (8), soviet (8). The added terms appear to be reasonable for this

$p(l_{ij})$	Holonyms	Hypernyms	Meronyms	Synonyms
0.1	0.1979	0.1963	0.1975	0.1982
0.2	0.1977	0.1941	0.1962	0.1984
0.3	0.1966	0.1898	0.1959	0.1961
0.4	0.1949	0.1810	0.1939	0.1925
0.5	0.1927	0.1716	0.1923	0.1885
0.6	0.1903	0.1595	0.1887	0.1798
0.7	0.1887	0.1454	0.1841	0.1706
0.8	0.1868	0.1278	0.1795	0.1635
0.9	0.1838	0.1145	0.1766	0.1513
1	0.1779	0.0938	0.1690	0.1402

Table 5.4: Retrieval effectiveness when incorporating different types of relationships

α	Qiu	PAS
0.25	0.2009	0.2000
0.5	0.2004	0.2012
0.75	0.2004	0.2015
1	0.1982	0.1997
2	0.1932	0.1877

Table 5.5: Mean average precision with Qiu and Frei's query expansion technique and PAS technique, for different values of α.

type of collection and requests. However, there is no significant increase of retrieval effectiveness for both techniques, which is quite disappointing compared to the 20-30% improvements obtained by Qiu and Frei on three test collections. One possible explanation is that the baseline is much higher than with the $tf-idf$ weighting scheme which was used in their experiments.

5.4.4 Determining relevant terms

To compute the degree of support of each term, we need two probability estimates: $p(a_i)$, in $a_i \rightarrow T_i$, and $p(l_{ij})$, in $T_i \wedge l_{ij} \rightarrow T_j$. For computing $p(a_i)$, Table 5.6 shows that as the number of occurrences qf of a query term in the query increases, its estimated probability of relevance (according to the precision test described in Section

qf	# of terms	# of rel. terms	Fraction of rel. terms
1	1413	702	0.497
2	55	38	0.69
3	5	4	0.8

Table 5.6: Relevance of term given their query frequency

β	PAS	IP
Baseline	0.1995	0.1971
0.25	0.2011	0.1982
0.5	0.2017	0.1993
0.75	0.2020	0.2002
1	0.2027	0.2011
1.5	0.2037	0.2024
2	0.2046	0.2032
2.5	0.2039	0.2031
3	0.2030	0.2031
4	0.1994	0.2023
5	0.1978	0.1986
6	0.1954	0.2005
8	0.1918	0.1986
10	0.1843	0.1975
12	0.1787	0.1968
14	0.1726	0.1959

Table 5.7: Mean average precision found for different β values, applied to the PAS and internal product (IP) ranking formula

5.2.2) to the information need increases. For example, when a term occurs twice, there is probability of 0.69 that this term will be useful for retrieval, while this probability is only 0.497 for one occurrence. We will take the frequency estimates shown in that table for $p(a_i)$.

We must also estimate $p(r_{ij})$ from the similarity $sim(t_i, t_j)$. We assumed that $p(l_{ij}) = \alpha \cdot sim(t_i, t_j)$. Our goal is to have the best possible estimate of the probability that a query term is relevant. Since it is possible to "know" which terms are relevant with the retrieval effectiveness test (section 5.2), we may attempt to find the value of α which will lead to the best estimate of this probability. For a given value of α, we computed the mean square error MSE between the estimated probabilities and the real probability (0:not relevant, or 1:relevant). We observed a monotonic decrease of the MSE for values of α from 0 to 6, then a monotic increase from $\alpha = 6$ to $\alpha = 20$. A good estimate is then: $\alpha = 6$. Once the degrees of support are computed, we can use Eq. 5.9 to find a new value for each term's weight, hopefully reflecting better the term's relevance to the information need. We have tested a series of value for the parameter β, as shown on table 5.7. We achieved 2.55% and 3.05% improvement of mean average precision for respectively the PAS and the internal product weighting functions, with $\beta = 2$. This improvement is not significant, but encouraging. Much work has to be made to find more refined probability estimates.

5.5 Discussion

In this chapter, we have developed a new technique for improving a query term weight using statistical similarities between terms, and demonstrated that the PAS logical model can be implemented if some approximations are allowed on the computation

of the degree of support. In the next chapter, we will discuss issues concerning these experiments with the PAS logical model.

Chapter 6

Conclusion

This conclusion will first make a review of the main contributions of this thesis (Section 6.1). The questions that remain opened or that were opened by this work will then be addressed (Section 6.2). A discussion on the future of the logical approach will conclude this thesis (Section 6.3).

6.1 Main contributions

This section makes a short comeback on the main contributions of this thesis:

A logical model based on propositional logic. There is a certain contradiction in stating on one side that retrieval is inference and that it should be treated with logic, and on the other side to affirm that propositional logic, the simplest and most intuitive of all logics, is inherently inadequate for that purpose. Sebastiani already made the demonstration that a misleading interpretation of the logical implication led to that belief, but he admitted the difficulty of of incorporating uncertainty inside propositional logic [Seb98]. With PAS, uncertainty is inherently part of the knowledge and of the inference processes, to the contrary of approaches which artificially graft uncertainty on top of logic. The PAS model is remarkable because the retrieval task is completely transparent, and the consequences of using logic are very clear.

The Logical Uncertainty Principle. Since it was proposed in 1986, the Logical Uncertainty Principle has been shown to be implicit to many retrieval models, and has influenced a number of works in the logical approach. The principle has been reinterpreted [Nie89], but it is the first time it is extended to encompass symbolic representations of information. This reinforces the idea that retrieval should be considered first as a symbolic process where each decision can be ”explained”, and that the numerical evaluation of uncertainty should come in second place. The demonstration that the Logical Uncertainty Principle is implicit to the symbolic support is also an interesting result for the theory of PAS, for which this interpretation of the symbolic support has never been proposed. Finally, this result links different views on the notion of partial entailment.

Logical treatment of hypertext retrieval. With the abundance of hypertextual information on the Web, a plethora of techniques and ideas have been proposed for

dealing with this type of information. At some point, a unifying formalism becomes necessary to understand and compare the different techniques to treat this information. To represent and process the knowledge induced by the the Web structure, artificial intelligence techniques attract much interest. It has been shown that Probabilistic Argumentation Systems propose a logical interpretation to the ideas behind four of the most important approaches to process hypertextual evidence - enhancing document content, spreading activation, measuring popularity, and finding hubs and authorities.

The Cluster Hypothesis for query terms. Weighting schemes generally ignore the semantic content of query terms, and their potential relevance to the information need of the user. The only ”semantic” processing consists in removing the stopwords, which is equivalent to setting their weight to a null value. It was shown in this thesis that term similarities can be used to help determining the relevant terms of the query and hence modify their weights. The experimental results were encouraging, but the technique must be refined in order to achieve significant improvements. It is our belief that with a more appropriate computation and use of term similarities, significant improvement can be achieved. Moreover, this Cluster Hypothesis could also be applied to improve the document weighting. A document may contain hundreds of words, while many of them might not be relevant to the topic addressed by this document. The large number of term similarities could help determining with more reliability than for queries the subset of relevant terms. The computational cost is not a problem, since all processing can be done before operating the retrieval system. In short, we believe that the Cluster Hypothesis for (query/document) terms should be further investigated.

Implementation on a large collection. To convince the whole IR community of their worth, logical models should be scalable to large collections which are now the standards for the IR community. This thesis has demonstrated that if reasonably simplifying assumptions are made, making logical inferences does not necessarily imply a high computational cost. The hypertext retrieval model which has been applied to a collection of 2.3 GB can be rather easily scaled up to a collection of more than a hundred GB since the cost is linear with the number of documents. It has also been shown that logic does not have to be the cornerstone of the retrieval system: it can be helpful as a component adequate for making logical processing of information, for example to determine the content-bearing terms of a query. This encourages the spreading of logic throughout the IR community, already pushed by new types of applications such as Question Answering systems, for which the need of logic is more apparent than for standard IR.

6.2 Open questions

There is a number of questions left opened by this work. Some of them are not new to IR, but it is interesting to interpret them within the PAS approach.

Obtaining probabilities. The problem of estimating accurately uncertainty is central to IR. Important advances have been made in this area, essentially within the probabilistic approach (see e.g. [RSJ76, Gre98]). PAS leave total freedom on the way probabilities can be assigned to assumptions, but freedom is not always

an advantage: some constraints or a general criteria would be helpful to determine adequate uncertainty values. Making inferences can be helpful if adequate uncertainty values are associated, but may be very harmful otherwise. Indeed, the problem of estimating probability is crucial to the whole logical approach, although logic appears as an attempt to move away from purely numerical approaches. In this thesis, different approaches were take to estimate probabilities of assumptions: they were considered as parameters which are optimized according to the retrieval effectiveness criteria (Chapter 5), they were adjusted assuming they follow a logistic regression on the rank (a priori assumptions, Chapter 4), or estimated using past relevance judgements (link assumptions, Chapter 4). However, a general methodology would have been much helpful.

Logic and empirical observations. The sources of evidence which may influence relevance are innumerable, and interact in complex and often misunderstood ways. Logic is used to represent and process this knowledge (e.g. relationships between concepts or documents), but can logic really capture all the subtle interactions between the parameters which influence retrieval? Does evidence combine and propagate in the way we suppose or assume they do? An earlier paper [Pic98] showed that for the CACM collection, hypertextual evidence appear to combine in the noisy-OR way, which is the combining scheme most naturally used within the PAS framework. However, this result was not confirmed on the WT collection, where hypertextual evidence are highly correlated and simplifying independence assumptions assumption cannot be made. It is then highly dangerous to assume without verification that evidence should behave in a certain way.

There is a more fundamental question: to what degree do the rules of inference apply to information retrieval? The point here is not whether IR is or is not inference, but whether there is strong empirical evidence that the relationships between the objects handled by information retrieval systems (e.g. terms which are indicators of the presence of concepts) follow the rules of inference. If very low associated probabilities are associated to rules (because higher probability values would harm retrieval), making inferences loses its meaning. If documents are represented by a large number of terms, or if they are linked to many documents, the inferences get "diluted" and the overall result of the inference process can get unpredictable and unreadable.

Knowledge representation. There is an increasing interest for NLP and more elaborated knowledge representation techniques, due to new applications such as Question Answering systems or more complex retrieval environments. But propositional logic is only able to handle facts, which may not be sufficient for many applications related to IR. For example, it cannot deal adequately with a question such as: "What Arab country invaded Kuwait in 1990?". Such a question can be naturally represented in first-order logic by:

$$\exists x,\ Arab(x) \wedge Invade(Kuwait, x, 1990)$$

Can PAS be built on more elaborated techniques such as first-order logic, instead of propositional logic? The basic ideas of PAS of representing uncertainty explicitly and viewing the evaluation of hypothesis as a research of arguments is very attractive for reasoning under uncertainty. PAS could be a very convenient framework to unify probabilistic knowledge with first-order logic in the context of the theory of evidence.

The computational cost of inferences. We have seen in Chapter 5 that a rigorous computation of the symbolic support ($sp(D, \xi \wedge Q)$) would have implied a cost per query linear with the number of documents. This is simply not possible with the ever increasing size of document collections. This leads to the following fundamental questions: are logical models destinated to be no more than a convenient framework for thinking of the IR task? Should logical processing be avoided as much as possible when moving from models to their implementation, or eventually only kept for specific processing? The work of Ounis and Pasca on image retrieval [OP98] is a negative answer to those questions. They showed that a logical representation of images outperforms a keyword-based representation, and found an efficient technique for resolution. This work should hopefully be followed by others in the following years. However, general purpose collections are in general difficult to adapt to a logical modeling.

6.3 The future of logical approaches to IR

At present time, building a retrieval system that yields good retrieval results holds more of a "know how" than of the rigorous application of general theories of text representation, user modeling or information retrieval. Indeed, IR is a highly diverse and complex task and as such, is quite refractory to modeling. Nevertheless, we have seen that models, even as naive approximations of reality, have been essential to the progresses done in IR in the last thirty years. They should become more and more essential as the complexity and diversity of text collections and information needs increases.

It is often said that relevance is the central concept of IR. In our opinion, relevance is "only" an emanation of a more general concept, the concept of information. To the author's advice, the fundamental questions of IR turn around the concept of information: how can the information contained in any object be represented explicitly? How can an information need (sometimes called "information gap"), which cannot be said to "contain" information, be represented inside a theory for representing information? What is the true meaning of "matching" a content of information with a gap of information, and how is it exactly captured by logical implication? The logical approach starts from an ambition to answer these questions. But while logic is a language to describe the state of the world, information is a process which involves a change in a cognitive agent receiving the data: it is not so obvious how this process can be explicitly described inside logic, although logic appears to be the most promising formalism to handle the notion of information [Dev91, vR96]. To elucidate the relationships between IR and logic, a good starting point is propositional logic, the simplest and best understood logic - although it was misunderstood in its application to IR. Modeling IR with probabilistic argumentation systems is one small step toward this far away goal.

Bibliography

[AHKL97] B. Anrig, R. Haenni, J. Kohlas, and N. Lehmann. Assumption-based modeling using ABEL. In D. Gabbay, R. Kruse, A. Nonnengart, and H.J. Ohlbach, editors, *First International Joint Conference on Qualitative and Quantitative Practical Reasoning; ECSQARU–FAPR'97*. Springer, 1997.

[BH94] P.D. Bruza and T.W. Huibers. Investigating aboutness axioms using information fields. In *Proc. of the Int. ACM-SIGIR Conf.*, pages 112–121, 1994.

[Bh98] K. Bharat and M. henzinger. Improved algoritms for topic distillation in hyperlinked environments. In *Proc. of the Int. ACM-SIGIR Conf.*, pages 104–111, 1998.

[BKFS95] N.J. Belkin, P. Kantor, E.A. Fox, and J.A. Shaw. Combining the evidence of multiple query representations for information retrieval. *Information Processing & Management*, 31(3):431–448, 1995.

[BL98] S. Brin and L.Page. The anatomy of a large-scale hypertextual web search engine. In *Proceedings of the World Wide Web Conference*, pages 107–117, 1998.

[BS74] A. Bookstein and D.R. Swanson. Probabilistic models for automatic indexing. *Journal of the American Society for Information Science*, 25:312–318, 1974.

[CC92] Y. Chiamarella and Y. Chevallet. About retrieval models and logic. *The Computer Journal*, 5(3):233–242, 1992.

[CCH92] J.P. Callan, W.B. Croft, and S.M. Harding. The INQUERY retrieval system. In *Proceedings of the 3rd International Conference on Database and Expert Systems Applications*, pages 78–83, 1992.

[CdBD99] S. Chakrabati, M. Van der Berg, and B. Dom. Focused crawling:A new approach to topic specific resource discovery. In *Proceedings of the World Wide Web Conference*, pages 545–567, 1999.

[CDR+98] S. Chakrabati, B. Dom, P. Raghavan, S. Rajagopalan, D. Gibson, and J. Kleiberg. Automatic resource compilation by analyzing hyperlink structure and associated text. In *Proceedings of the World Wide Web Conference*, 1998.

[CH79] W.B. Croft and D.J. Harper. using probabilistic models of document retrieval without relevance information. *Journal of Documentation*, 35(4):285–295, 1979.

[Cle84] C.W. Cleverdon. Optimizing convenient on-line access to bibliographic databases. *Information Service & Use*, 4:37–47, 1984.

[CMK66] C.W. Cleverdon, J. Mills, and M. Keen. Factors determining the performance of indexing systems. ASLIB Cranfield Research Project. Technical report, Cranfield, UK, 1966.

[CRSvR96] F. Crestani, I. Ruthven, M. Sanderson, and C.J. van Rijsbergen. The troubles with using a logical model of IR on a large collection of documents. *TREC-4*, pages 509–526, 1996.

[CS00] A. Le Calvé and J. Savoy. Database merging strategy based on logistic regression. *Information Processing & Management*, 2000. To appear.

[CT87] W.B. Croft and R.H. Thompson. I3R: A new approach to the design of document retrieval systems. *Journal of the American Society for Information Science*, 38(6):389–404, 1987.

[CT93] W.B. Croft and H.R. Turtle. Retrieval strategies for hypertext. *Information Processing & Management*, 29(3):313–324, 1993.

[CvR95] F. Crestani and C.J. van Rijsbergen. Information retrieval by logical imaging. *Journal of Documentation*, 51(1):3–17, march 1995.

[Dev91] K.J. Devlin. *Logic and Information*. Cambridge University Press, Cambridge, England, 1991.

[dK86] J. de Kleer. An assumption-based tms. *Journal of Artificial Intelligence*, 28:127–162, 1986.

[FB89] N. Fuhr and C. Buckley. Optimum retrieval functions based on the probability ranking principle. *ACM Transactions on Information Systems*, 7(3):183–204, 1989.

[FGR98] N. Fuhr, N. Govert, and T. Rolleke. DOLORES: A system for logic-based retrieval of multimedia objects. In *Proc. of the Int. ACM-SIGIR Conf.*, pages 257–265, 1998.

[FNL88] E.A. Fox, G.L. Nunn, and W.C. Lee. Coefficients for combining concept classes in a collection. In *Proc. of the Int. ACM-SIGIR Conf.*, pages 291–307, 1988.

[FS95] H.P. Frei and D. Stieger. The use of semantic links in hypertext information retrieval. *Information Processing & Management*, 31(1):1–13, 1995.

[Fuh95] N. Fuhr. Probabilistic Datalog- A logic for powerful retrieval models. In *Proc. of the Int. ACM-SIGIR Conf.*, pages 282–290, 1995.

[Fuh00] N. Fuhr. Probabilistic Datalog Implementing logical information retrieval for advanced applications. *Information Processing & Management*, 51(2):95–110, 2000.

[Gey94] F.C. Gey. Inferring probability of relevance using the method of logistic regression. In *Proc. of the Int. ACM-SIGIR Conf.*, pages 222–231, 1994.

[Gre98] W.R. Greiff. A theory of term weighting based on exploratory data analysis. In *Proc. of the Int. ACM-SIGIR Conf.*, pages 11–19, 1998.

[Hei89] K.D. Heidtmann. Smaller sums of disjoint products by subproduct inversion. *IEEE Transactions on Reliability*, 38(3):305–311, 1989.

[HKL99] R. Haenni, J. Kohlas, and N. Lehmann. Probabilistic argumentation systems. Technical Report 99-09, Institute of Informatics, University of Fribourg, 1999.

[HL98] R. Haenni and N. Lehmann. Reasoning with finite set constraints. In *ECAI'98, Workshop W17: Many-valued logic for AI application*, pages 1–6, 1998.

[KC99] M.C. Kim and K.S. Choi. A comparison of colocation-based similarity measures in query expansion. *Information Processing & Management*, 35(1):19–30, 1999.

[KH96] J. Kohlas and R. Haenni. Assumption-based reasoning and probabilistic argumentation systems. In J. Kohlas and S. Moral, editors, *Defeasible Reasoning and Uncertainty Management Systems: Algorithms*. Oxford University Press, 1996.

[Kle98] J. Kleinberg. Authoritative sources in a hyperlinked environment. In *Proceedings of the 9th Symposium on Discrete Algorithms*, pages 668–677, 1998.

[Lal97] M. Lalmas. Dempster-shafer's theory of evidence applied to structured documents. In *Proc. of the Int. ACM-SIGIR Conf.*, pages 110–118, 1997.

[Lal98] M. Lalmas. Logical models in information retrieval: Introduction and overview. *Information Processing & Management*, 34(1):19–33, 1998.

[LG99] S. Lawrence and C.L. Giles. Accessibility of information on the web. *Nature*, 400(8):107–109, 1999.

[LL89] K.B. Laskey and P.E. Lehner. Assumptions, beliefs and probabilitie. *Artificial Intelligence*, 41:67–77, 1989.

[LR98] M. Lalmas and I. Ruthven. Representing and retrieving structured documents using the Dempster-Shafer theory of evidence: modeling and evaluation. *Journal of Documentation*, 54(5), 1998.

[Luh57] H.P. Luhn. A statistical approach to the mechanized encoding and searching of litterary information. *IBM Journal of Research and Devlopment*, 1:309–317, 1957.

[Mar64] J. Martyn. Bibliographic coupling. *Journal of Documentation*, 20(4):236, 1964.

[Mar97] M. Marchiori. The quest for coorect information on the Web:Hyper search engines. In *Proceedings of the World Wide Web Conference*, 1997.

[Miz97] S. Mizzaro. Relevance:The whole history. *Journal of the American Society for Information Science*, 48(9):810–832, september 1997.

[MK60] M.E. Maron and J.L. Kuhns. On relevance, probabilistic indexing and information retrieval. *Journal of the ACM*, 7(23):1216–244, 1960.

[Nau70] D. Nauta. *The meaning of information*. Mouton, The Hague (NL), 1970.

[Nie89] J. Nie. An information retrieval model based on modal logic. *Information Processing & Management*, 25(5):477–491, 1989.

[Nie96] J. Nie. An inferential approach to information retrieval and its implementation using a manual thesaurus. *Artificial Intelligence Review*, 10:409–439, 1996.

[OP98] I. Ounis and M. Pasca. RELIEF: Combing expressiveness and rapidity into a single system. In *Proc. of the Int. ACM-SIGIR Conf.*, pages 266–274, 1998.

[otCP99] Members of the Clever Project. Hypersearching the web. *Scientific American*, 1999.

[PC98] J.M. Ponte and W.B. Croft. A language modeling approach to information retrieval. In *Proc. of the Int. ACM-SIGIR Conf.*, pages 275–281, 1998.

[Pic98] J. Picard. Modeling and combining evidence provided by document relationships using probabilistic argumentation systems. In *Proceedings of the International ACM-SIGIR Conference on Research and Development in Information Retrieval*, pages 182–189, Melbourne, Australia, 1998.

[Pic99] J. Picard. Finding content-bearing terms using term similarities. In *Ninth Conference of the European Chapter of the Association for Computational Linguistics*, pages 241–244, Bergen, Norway, 1999. Student session.

[PW91] H.J. Peat and P. Willet. The limitations of term co-occurence data for query expansion in document retrieval systems. *Journal of the American Society for Information Science*, pages 378–383, June 1991.

[QF93] Y. Qiu and H.P. Frei. Concept based query expansion. In *Proc. of the Int. ACM-SIGIR Conf.*, pages 160–169, 1993.

[RF98] T. Rolleke and N. Fuhr. Information retrieval with probabilistic Datalog. In *Information retrieval: Uncertainty and logic*, chapter 9, pages 221–243. Kluwer, 1998.

[Rob77] S.E. Robertson. The probability ranking principle in information retrieval. *Journal of Documentation*, 33(4):294–304, 1977.

[RRS98] A. Rai, T. Ravichadan, and S. Samaddar. How to anticipate the Internet's global diffusion. *Communications of the ACM*, 41(10):97–106, 1998.

[RSJ76] S.E. Robertson and K. Sparck-Jones. Relevance weighting of search terms. *Journal of the American Society for Information Science*, 27(3):129–146, 1976.

[RTT99] R.Mandala, T. Tokunaga, and H. Tanaka. Combining general hand-made and automatically constructed thesauri for information retrieval. In *IJCAI'99*, pages 920–925, 1999.

[RW94] S.E. Robertson and S. Walker. Some simple effective approximations to the 2-Poisson model for probabilistic weighted retrieval. In *Proc. of the Int. ACM-SIGIR Conf.*, pages 232–241, 1994.

[RW97] S.E. Robertson and S. Walker. On relevance weights with little relevance information. In *Proc. of the Int. ACM-SIGIR Conf.*, pages 16–24, 1997.

[RWHB95] S.E. Robertson, S. Walker, and M.M. Hancock-Beaulieu. Large test collection experiments on an operational, interactive system: OKAPI at TREC. *Information Processing & Management*, 31(3):345–360, 1995.

[Sav94] J. Savoy. A learning scheme for information retrieval in hypertext. *Information Processing & Management*, 30(4):513–533, 1994.

[Sav95] J. Savoy. A new probabilistic scheme for information retrieval in hypertext. *The new Review of Hypermedia and Multimedia*, pages 107–131, 1995.

[Sav97] J. Savoy. Ranking schemes in hybrid Boolean systems: A new approach. *Journal of the American Society for Information Science*, 48(3):235–253, 1997.

[Seb98] F. Sebastiani. On the role of logic in information retrieval. *Information Processing and Management*, 34(1):1–18, 1998.

[SJ72] K. Sparck-Jones. A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 28:11–21, 1972.

[SK93] P. Smets and R. Kruse. The transferable belief model for belief representationy. In A. Motro and P. Smets, editors, *Uncertainty Management in information systems: from needs to solutions*, pages 343–368. Kluwer, 1993.

[SM83] G. Salton and M.J. McGill. *The SMART and SIRE experimental retrieval systems*. McGraw-Hill, New York, 1983.

[Sma93] F. Smadja. Retrieving collocations from text: Xtract. *Computational linguistic*, 19(1):143–177, 1993.

[SP97] H. Schutze and J.O. Pedersen. A cooccurrence-based thesaurus and two applications to information retrieval. *Information Processing & Management*, 33(3):307–318, 1997.

[SvR83] A.F. Smeaton and C.J. van Rijsbergen. The retrieval effects of query expansion on a feedback document retrieval system. *The Computer Journal*, 26(3):239–246, 1983.

[SWY75] G. Salton, A. Wong, and C.S. Yang. A vector space model for automatic indexing. *Communications of the ACM*, 18:613–620, 1975.

[TC91] H. Turtle and W.B. Croft. Evaluation of an inference network-based retrieval model. *ACM Transactions on Information Systems*, 9(3):187–222, 1991.

[TC92] H. Turtle and W.B. Croft. A comparison of text retrieval models. *The Computer Journal*, 35(3):279–290, 1992.

[TC97] H. Turtle and W.B. Croft. Uncertainty in information retrieval systems. In A. Motro and P. Smets, editors, *Uncertainty management in information system*, Amsterdam (NL), 1997. Kluwer.

[Voo93] E.M. Voorhees. Using wordnet to disambiguate word senses for text retrieval. In *Proc. of the Int. ACM-SIGIR Conf.*, 1993.

[Voo94] E.M. Voorhees. Query expansion using lexical semantic relations. In *Proc. of the Int. ACM-SIGIR Conf.*, 1994.

[vR77] C.J. van Rijsbergen. A theoretical basis for the use of co-occurrence data in information retrieval. *Journal of Documentation*, 33(2):106–119, 1977.

[vR79] C.J. van Rijsbergen. *Information retrieval*. Butterworths, London, UK, 2nd edition, 1979.

[vR86] C.J. van Rijsbergen. A non classical logic for information retrieval. *Journal of Documentation*, 29(6):481–485, 1986.

[vR89] C.J. van Rijsbergen. Towards an information logic. In *Proc. of the Int. ACM-SIGIR Conf.*, pages 77–86, 1989.

[vR96] C.J. van Rijsbergen. Information retrieval and informative reasonning. In J. van Leeuven, editor, *Lecture Notes in Computer Science #1000*, pages 549–559. Springer Verlag, 1996.

[vRL96] C.J. van Rijsbergen and M. Lalmas. Information calculus for information retrieval. *Journal of the American Society for Information Science*, 47(5):385–398, 1996.

[vRSJ73] C.J. van Rijsbergen and K. Sparck-Jones. A test for the separation of relevant and non-relevant documents in experimental retrieval collections. *Journal of Documentation*, 29(3):251–257, September 1973.

[XC96] J. Xu and W.B. Croft. Query expansion using local and global document analysis. In *Proc. of the Int. ACM-SIGIR Conf.*, pages 4–11, 1996.

[Zip49] G.K. Zipf. *Human behaviour and the principle of least effort*. Adison-Wesley, Reading (MA), 1949.