The Troubles with Using a Logical Model of IR on a Large Collection of Documents

Experimenting Retrieval by Logical Imaging on TREC

F. Crestani†, I. Ruthven*‡, M. Sanderson‡, C.J. van Rijsbergen‡

†Dip. di Elettronica e Informatica - Universitá di Padova - Italy
‡Dept. of Computing Science - University of Glasgow - Scotland

Abstract

The evaluation of an implication by Imaging is a logical technique developed in the framework of Conditional Logics. In 1993 a logical model of IR called "Retrieval by Logical Imaging" was proposed by some of the authors of this paper and tested using some classical IR test collections.

In this paper we report on the challenges posed by trying to apply such a model to a large test collection of the size of TREC-B. The problems we found and the way we put together ideas and efforts to solve them are indicative of the troubles one might find in trying to implement and experiment with a "complex" logical model of IR. We believe our efforts could set an example for other researchers working on logical models of IR to try to implement their models in such a way that they can cope with the size of real life collections, though preserving the formal "beauty" of their logical models.

*Address to which correspondence should be sent: Ian Ruthven, Department of Computing Science, University of Glasgow, Glasgow G12 8QQ, Scotland, UK; email: trec@dcs.gla.ac.uk

1 The use of non-classical logic in Information Retrieval

In recent years there have been several attempts to define *a logic for IR*. The earliest approaches were directed to the use of classical logic, like Boolean logic [Sal68], with few notable exceptions [Hil65]. The basis of a logical model for IR is the assumption that queries and documents can be represented by logical formulas. In order to retrieve a document an IR system has to infer the query from formulas representing the document. This logical interpretation of query and documents emphasises that *IR is an inference process* by which we can infer if a document is relevant to the query using information present in the document itself together with user knowledge. In classical logic inference is often associated with *logical implication*: a document is relevant to a query if it implies the query, that is if $d \rightarrow q$ is true. In IR it was soon realised that it is necessary to take into consideration the uncertainty inherent in this implication. A collection of documents cannot be considered as a consistent set of statements, and documents in the collection could even contradict each other. In order to cope with uncertainty a logic for probabilistic inference was introduced. If $d \rightarrow q$ is uncertain, then we can measure its degree of uncertainty by $P(d \rightarrow q)$. An early suggestion was to estimate $P(d \rightarrow q)$ by $P(q/d)$. However, the limitation of this approach explained by the triviality results of Lewis [Lew81] excluded that conditional probabilities could be used as probabilistic logic dealing with conditionals.

In 1986 Van Rijsbergen [vR86] proposed the use of a non-classical conditional logic for IR. The proposal initiated a new line of research that was followed by many researchers (see for example [Nie88, Nie89, CC92, Bru93]).

A few years later Van Rijsbergen proposed to estimate the probability of the conditional by a process called Imaging [vR89]. This idea was finally put into an implementation in 1994 when Crestani and Van Rijsbergen [CvR95] proposed a retrieval technique called *Retrieval by Logical Imaging* (RbLI). This technique enables the evaluation of $P(d \rightarrow q)$ and $P(q \rightarrow d)$ by Imaging according to a "Possible Worlds" semantics where a term is considered as a possible world. This technique exploits term-term relationships in retrieval by means of an accessibility relation between worlds based on their Expected Mutual Information Measure (EMIM).

This paper reports on the problems, solutions, and current results of the experimentation of Retrieval by Logical Imaging using a large collection of documents. The paper is structured as follows. In Section 2 we present the model, while in Section 3 we lay out the experimental settings for the implementation of the model. This is where the problems start. Experimenting with a large collection, of the size of TREC-B, poses considerable difficulties that are reported in Section 4. Section 5 describes our attempted solutions towards an implementation of the RbLI model that could cope with the size of the test collection. Section 6 reports our current results in the context of the TREC-4 initiative, "ad hoc" track. Further directions of investigation are described in Section 7.

2 Retrieval by Logical Imaging

Logical Imaging (LI) is a logical technique that enables the evaluation of a conditional sentence without explicitly defining the operator "$\rightarrow$" [Sta81]. LI is based on the Possible World Semantics [Kri71], that is on a semantics where the truth value of a logical sentence is evaluated in the context of a "world", or as Hughes called it in the context of a "conceivable or envisageable state of affairs" ([HC68], p. 75). According to this semantics the truth value of the conditional $y \rightarrow x$ in a world w is equivalent to the truth value of the consequent x in the closest world w_y to w where the antecedent y is true.

LI was extended to the case where there is a probability distribution on the worlds by Lewis [Lew81]. In this case the evaluation of $P(y \rightarrow x)$ causes a shift of the original probability P from a world w to the closest world w_y where y is true. Probability is neither created nor destroyed, it is simply moved from a "not-y-world" to a "y-world" to derive a new probability distribution P_y. This process is called "deriving P_y from P by imaging on y".

We will not go here into a detailed explanation of the Imaging process. The interested reader can look at papers by Stalnaker [Sta81], Lewis [Lew81], and Gärdenfors [G88] for more details.

We use Imaging in IR with the purpose of estimating the probability of relevance of a document by means of the probability of the conditional $d \rightarrow q$, by assuming that:

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

We call *Retrieval by Logical Imaging* (RbLI) a model that produces a ranking of every document d_i in the collection based on an estimate of $P(d_i \rightarrow q)$. A detailed explanation of this model can be found in [CvR95]. Briefly, in RbLI we derive P_d by imaging on d considering all the index terms t in T as possible worlds. If so, $P(d \rightarrow q)$ can be evaluated as:

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

where $P(t)$ and $P_d(t)$ are respectively the "prior" and the "posterior" probability assigned on the space T, t_d is the closest term to t for which d is true, or in other words, the most similar term to t that occurs in the document d, and $I(t_d, q)$ is defined as:

$$I(t_d, q) = \begin{cases} 1 & \text{if } t_d \in q \\ 0 & \text{otherwise} \end{cases}$$

The application of the above technique to IR requires a probability distribution on T so that for every t we can have $P(t)$, and a measure of similarity over the term space T to enable the identification of t_d. We will tackle this problem in Section 3.

LI provides the minimal revision of the "prior" probability in the sense that it involves no gratuitous movement of probability from world to dissimilar worlds. In fact, the revision of the "prior" probability necessary to make d certain is obtained by adopting the least drastic change in the probability space. This is achieved by transferring probabilities from each term not occuring in the document d to its closest (the most similar) term occurring in it, so that the total amount of the distance covered in the transfer is minimal. A detailed comparison between conditional probability and the conditionalisation performed by Imaging can be found in [Cro94].

For a practical example of the evaluation of RbLI let us suppose we have a document d represented by terms t_1, t_5, and t_6 and a query q represented by

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

Table 1: Evaluation of $P(d \rightarrow q)$ by imaging on d

t_1, t_4, and t_6. Each of these terms has a "prior" probability associated with it, this is indicated by $P(t)$. Table 1 reports the evaluation of $P(d \rightarrow q)$ by imaging on d. The evaluation process is the following:

1. Identify the terms occurring in the document d (third column of the table).
2. Determine for each term in T the t_d, i.e. the most similar term to t for which $I(t, d) = 1$. This is done using a similarity measure on the term space (fourth column).
3. Evaluate $P_d(t)$ by transferring the probabilities from terms not occurring in the document to terms occurring in it (fifth column).
4. Evaluate $I(t, q)$ for each term, i.e. identify the terms occurring in the query (sixth column).
5. Evaluate $P_d(t) \cdot I(t, q)$ for all terms (seventh column) and evaluate $P_d(q)$ by summation (bottom of seventh column).

A graphical interpretation of this process is depicted in Figure 1.

3 Implementing Retrieval by Logical Imaging

As it has been introduced in the previous section, in order to implement the RbLI model we require:

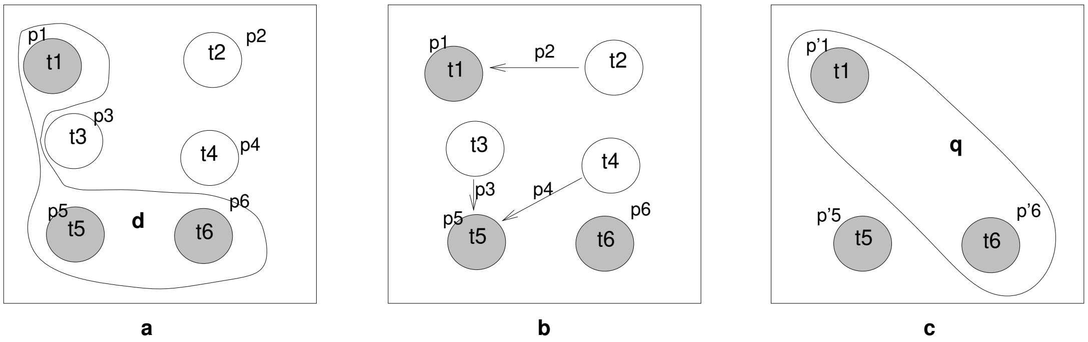

Figure 1: Graphical interpretation of the evaluation of $P(d \to q)$ by imaging on d.

1. a "prior" probability distribution over the index term space that should reflect the importance of each index term in the term space;
2. a measure of similarity (or alternatively a distance) between index terms;

These two requirements reflects the use of a Possible World Semantics, since they correspond to the probability distribution, and to the accessibility relation measure among the possible worlds [Lew81].

The problem of determining an appropriate "prior" probability distribution over the set of terms used to index a document collection is one of the oldest problems in IR and many models have been proposed for this purpose. The problem could be translated into finding a so called "measure of the importance of the term in the term space". In IR several discrimination measures have been proposed (see for example [vR79, RS76]) and there it is not clear which one should be prefered to the others. For the experiments performed in TREC we used the *Inverse Document Frequency* (*idf*) defined as:

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

where n is the number of documents in which the term t occurs, and N is the total number of documents in the collection.

Strictly speaking, this is not a probability measure since $\sum_t idf(t) \neq 1$, however since we assume it to be monotone to $P(t)$, we can use it instead of a proper probability function because we are only interested in a ranking of the documents of the collection, not in the exact probability values.

The problem of measuring the similarity between index terms in order to define a measure of accessibility among worlds is more difficult. It is very important to chose the appropriate measure since much of the power of RbLI depends on it. For our TREC experiments we used the *Expected Mutual Information Measure* (EMIM). The EMIM between two index terms is often interpreted as a measure of the statistical information contained in the first term about the other one (or vice versa, it being a symmetric measure). EMIM is defined as follows:

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

where i and j are binary variables representing terms.

When we apply this measure to binary variables we can estimate EMIM between two terms using the technique proposed in [vR77] p. 130. Using this measure we can evaluate for every term a ranking of all the other terms according to their decreasing level of similarity with it. We store this information in a file which is used at run-time to determine for an index term the most similar other index term that occurs in the document under consideration.

4 Experimenting using a large document collection

In [CvR95] the performance of the RbLI model was tested using a the *Cranfield* document collection. In [Cv95] the RbLI model was generalised into the RbGLI model and both these models were tested using the Cranfield, the *CACM*, and the *NPL* document collections. In this last paper RbLI and RbGLI were compared with two prototypical classical retrieval models and the following experimental result was found ([Cv95], p. 298):

> " . . . we believe we have shown that *in principle* a probability transfer that takes into account a measure of similarity between the donor and the recipient is more effective in the context of IR than a probability transfer that does not take that into account. Most current probabilistic retrieval models are based on a probability kinematics that does not take into account similarity between terms or between documents, unless "ad hoc" weighting schemas, mostly based on clustering, are used. We would therefore like to suggest a further investigation into more complex and optimised models for probabilistic retrieval, where probability kinematics follows a non-classical approach."

In order to ensure that this result does not depend on the small size of the document collections we used, we decided to test these models on a collection of much larger size. Moreover, we decided to compare their performance with that of real IR systems, ones that could be recognised as a tough "benchmark".

We decided to procede in two steps: first implement RbLI and test it, and only later implement RbGLI. The RbGLI model is more computationaly demanding due to the introduction of a "probability transfer function" that enables the probability to be tranferred not only to a single t_d but to a set of them according to their respective distance to the term under consideration.

The implementation and testing of RbLI in [CvR95] and [Cv95] were quite heavy due to inefficient implementations of the probability transfer and to the complexity of the models. These problems, that were easily solved on small document collections, are much more difficult to tackle with a large document collection. In the following section we report on the challenge posed by trying to make the RbLI work on a large document collection. Our participation in TREC-4 was in the smaller part B collection which consists of $165,000$ documents.

5 Getting RbLI to work

Because of the anticipated high computational load of performing RbLI on the TREC-B collection, it was decided to initially run experiments on a sub set of TREC-B so as to prototype the RbLI software being developed for these experiments. Rather than remove documents from the collection, it was decided to reduce all documents (which are in fact news articles) to just their lead paragraph, which generally for news articles is a summary of the article. This reduction resulted in a 70Mb document collection. All work reported in this section is based on this modified collection.

The implementations of RbLI reported in Crestani and van Rijsbergen [Cv95] were performed on small test collections. Because of their size, it was possible to compute in a reasonable time the probability transfer of all terms in each document in the collection. On the TREC-B collection however, it was calculated that to perform this complete transfer would take too long given current computing resources. So methods of optimising the probability transfer were investigated.

5.1 Reducing the number of transfers

The first area looked at was the accessibility relation used to determine how probabilities are transferred, namely the EMIM measure. One of the features of EMIM is its ability to compute the relatedness between any two terms even if those terms don't co-occur, which means that in the case of RbLI it would be possible to compute probability transfers between all terms. The EMIM measure calculated between terms that don't co-occur however was found to be close to a small constant value, so for the experiments reported in this section a decision was made to restrict the EMIM calculations to only those terms that co-occur. When transferring probabilities onto a document's terms, any term that doesn't co-occur with that document's terms will have its probabilities uniformly distributed to all those document terms, so as not to lose its probability.

By calculating the EMIM between only co-occurring terms, the total number of calculations to be performed is reduced by around 95%. But it was felt that this reduction could and should be improved with further optimisations.

5.2 More speed

Now that probability transfers were reduced to just those terms that co-occur, the speed of RbLI on a document collection becomes proportional to the number of term co-occurrences in that collection. Therefore if we want to speed up RbLI, we need to reduce the number of these co-occurrences. There are a several ways in which this might be done. For example one could choose to only use those term co-occurrences where the two co-occurring terms appear in the same paragraph. Indeed this possibility might be exploited in the future, for the time being however it was decided to investigate the speed increase on RbLI when whole terms are removed from the collection. This technique was already proposed and tested by Crestani and Van Rijsbergen in [CvR95], where an intuitive explanation of its usefulness was given. We wanted to test its effectiveness on a large document collection.

In choosing terms to be removed, the question arises which type of terms should we concentrate on: (a) the few terms that occur in many documents, or (b) the many terms that occur in a few documents, indeed is there any difference between the two? To answer this question, we need to examine the imaging process in more detail.

The part of RbLI that is the most computationally intensive is the final stage where probabilities are transferred onto the terms of each collection document. The time taken to complete this stage is proportional to the total number of probability transfers that will potentially be made. The term "potential" is used because not all transfers will happen. RbLI demands that even if a term co-occurs with several document terms, that term will only transfer its probability to just one of those document terms, its most similar. Nevertheless each one of these potential transfers has to be considered by the RbLI software, so each potential transfer does add to the time taken to complete this task. The number of potential transfers can be calculated using the following formula:

$$number\ of\ transfers = \sum_{d=1}^{D} \sum_{t=1}^{T_d} O_t$$

where: D is the set of all documents in the collection, T_d is the set of terms contained in document d, and O_t is the number of terms that co-occur with t.

The formula for O_t is as follows:

$$O_t = \sum_{d=1}^{D_t} (N_d - 1)$$

where: D_t is the set of documents containing term t, and N_d is the number of distinct terms in document d

So, for example, given the choice of, case A, removing 1 term that occurs in 100 documents or, case B, removing 50 terms that each occur in 2 documents, we can use these formula to calculate which of these choices will reduce the number of transfers the most. For example, if we assume that each document contains 10 distinct terms, then the reduction in the number of transfers resulting from the two term removal cases is as follows[1]:

For the first case

$$number\ of\ tranfers(A) = 100 * 1 * (100 * 9) = 90,000$$

For the second case

$$number\ of\ tranfers(B) = 50 * (2 * *1 * (2 * 9)) = 1,800$$

From this, we conclude that efforts should be concentrated on reducing the small number of terms that occur frequently in the collection.

5.3 Reducing the small number of terms that occur frequently in the collection

In initial experiments a standard stop list (taken from Van Rijsbergen [vR79] p. 18) was used when indexing TREC, but in the light of the work described above, it was decided to investigate how retrieval performance would be affected when a bigger stop list was used.

Using a standard $tf \cdot idf$ retrieval system, the effect on retrieval performance from using a number of different stop lists was tested. The definition of a term's membership for these stop lists was based on the number of documents that term occurred in. Stop lists were generated for terms that occurred in more than 90% of documents, more than 80% of documents,

[1] The removal of term(s) has another minor influence on the time taken to complete RbLI: all terms co-occuring with the term(s) being removed will have fewer probability transfers to them. The effect of this influence however is the same for both cases, and so it need not be considered.

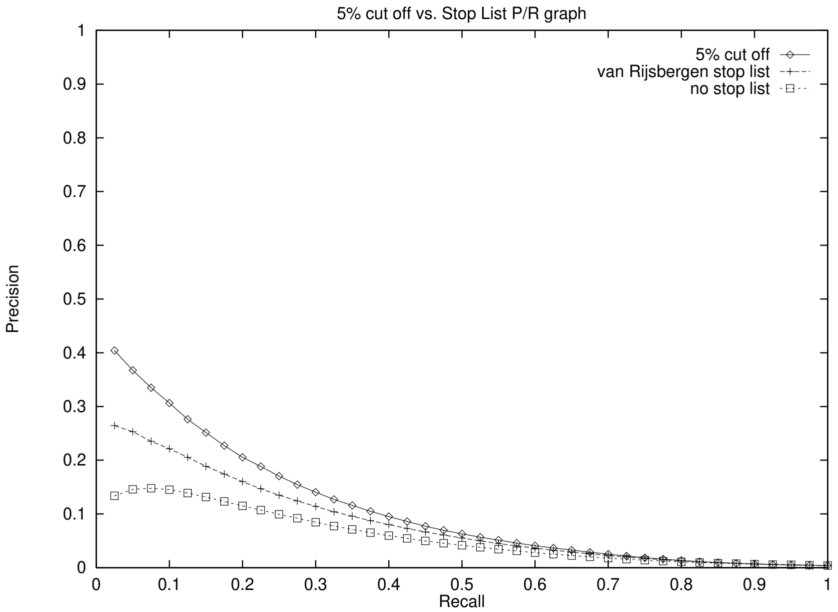

Figure 2: Precision and recall figures with different stop lists.

more than 70% of documents, and so on down to 2.5% of documents. For each stop list, the modified TREC-B collection was re-indexed, a retrieval run was performed and recall precision figures were obtained for each run. In addition two extra runs where performed where no stop list was used and a standard stop list from Van Rijsbergen [vR79] was used. The graph in figure 2 shows a selection of these runs.

As can be seem from the graph, precision is improved at all recall levels when a standard stop list is used instead of no stop list. If a stop list containing terms occurring in more than 5% of the documents is used however, there is a further uniform improvement in precision. The terms in this stop list account for around 50% of the term occurrences in the collection, as can be seen from Table 2. Although not plotted in this graph, it was found that using a larger stop list reduced precision.

From this experiment, it was concluded that the 5% stop list should be used when indexing the TREC-B collection so as to improve the speed of the RbLI process.

Initial number of occurrences	12,594,371
Occurrences after 5% stop list	6,902,676
Number of words in 5% stop list	253

Table 2: Effects of a 5% stop list

6 Evaluating Retrieval by Logical Imaging using the TREC-B document collection

Using the results reported in the previous section, we tried to perform a few experiments using the improved RbLI software on the *TREC-B document collection* for the *ad hoc* track. However, we soon run into a lot of small technical problems, the kinds of problems that almost all first time TREC participants experienced. Our lack of experience in dealing with large document collections together with the complexity of the RbLI model, made it impossible to have retrieval results ready for the TREC deadline.

In August 1995, when it became clear that the RbLI experiments would not be ready in time for the TREC-4 deadline, we decided to use a more classical IR system we already had developed in Glasgow to produce the retrieval results for the "ad hoc" track to send to TREC-4. We thought that we could later use the results of this system as a benchmark for the RbLI results, when these would be ready. The *glair* result set was then quickly generated using the system described in the following section and submitted instead of the RbLI results, that were not yet available.

6.1 The *glair* benchmark

The benchmark system we adopted for comparison with RbLI is "text book" IR system. It is based on the classical $tf \cdot idf$ retrieval strategy. Terms found in documents and queries have first their case normalised, then, any of these terms appearing in a stop list (taken from Van Rijsbergen [vR79] [2]) were removed. The remaining terms were suffix stripped using the Porter

[2]The stop list experiments reported in the previous section had not been carried out at this stage.

stemmer [Por80]. Document terms were weighted using the $tf \cdot idf$ weighting scheme. The *idf* formula has already been defined in Section 3, tf is defined as in [FB92]:

$$tf_{i,j} = \frac{\log(freq_{i,j} + 1)}{\log(length_j)}$$

where $freq_{i,j}$ is the frequency of term t_i in document d_j, and $length_j$ is the number of unique terms in document d_j.

The $tf \cdot idf$ retrieval strategy simply evaluates the product of the two components and ranks the documents in the collection based on a score. The score for each document is calculated by summing the $tf \cdot idf$ weights of any query terms found in that document.

We submitted to the TREC-4 Conference (TREC-B) the results of this system with low expectations. However, the results we achieved were not bad at all, as is summarised in the following excerpt from the official TREC-4 results:

```
Queryid (Num):       all  glair1
Total number of documents over all queries
    Retrieved:    49000
    Relevant:      2480
    Rel_ret:       1703
Interpolated Recall - Precision Averages:
    at 0.00       0.6321
    at 0.10       0.4332
    at 0.20       0.3085
    at 0.30       0.2507
    at 0.40       0.1747
    at 0.50       0.1453
    at 0.60       0.1035
    at 0.70       0.0744
    at 0.80       0.0543
    at 0.90       0.0393
    at 1.00       0.0229
Average precision (non-interpolated) over all rel docs
                  0.1819
Precision:
  At    5 docs:   0.3633
  At   10 docs:   0.3122
  At   15 docs:   0.2857
```

```
  At   20 docs:   0.2612
  At   30 docs:   0.2313
  At  100 docs:   0.1422
  At  200 docs:   0.0965
  At  500 docs:   0.0566
  At 1000 docs:   0.0348
R-Precision (precision after R (= num_rel for a query)
docs retrieved):
    Exact:        0.2170
```

The system, in the context of the TREC-B only participants, gave the best performances in 10 queries, and the worst performances in 6 queries out of the 49 used in TREC-B. Its overall performance was well above the median value of the average precision.

6.2 The RbLI results

Unfortunately, at the time of writing, we were still not able to obtain performance figures for RbLI. A few unexpected problems and some "bugs" in the RbLI software make impossible for us to have retrieval results from RbLI that can be compared with those obtained by our benchmark. Despite that, we think that this paper could provide a proof of our efforts toward using a complex logical model of IR on a large collection of documents, a task very rarely attempted by other researchers working on logical models for IR. We hope we set an example.

We will present the results of the use of RbLI on the TREC-B collection as soon as they are obtained.

7 Conclusions

We have been told by others that there is a tradition that "TREC first timers" fail to get their planned experiments done by the required deadline. We unfortunately have done nothing to change this.

Trying to implement RbLI on the TREC collection is proving to be a compromise between the theoretical purity demanded by the model, and the implementation problems posed by a collection of the size of TREC-B. We have found this compromise to a be a driving force in revealing other areas of work to be investigated. Therefore, experimental results aside, we regard our first participation in TREC as having been beneficial.

Acknowledgments

Most part of the work reported in this paper was supported by the ESPRIT Project *FERMI* (BRA 8134) on the "Formalisation and Experimentation of the Retrieval of Multimedia Information".

References

[Bru93] P.D. Bruza. *Stratified Information Disclosure: a synthesis between Hypermedia and Information Retrieval.* Phd thesis, Katholieke Universiteit Nijmegen, The netherlands, 1993.

[CC92] Y. Chiaramella and J.P. Chevallet. About retrieval models and logic. *The Computer Journal*, 35(3):233–242, 1992.

[Cro94] C.B. Cross. Eliminative bayesianism and probability revision. Unpublished paper, August 1994.

[Cv95] F. Crestani and C.J. van Rijsbergen. Probability kinematics in information retrieval. In *Proceedings of ACM SIGIR*, pages 291–299, Seattle, WA, USA, July 1995.

[CvR95] F. Crestani and C.J. van Rijsbergen. Information Retrieval by Logical Imaging. *Journal of Documentation*, 51(1):1–15, 1995.

[FB92] W.B. Frakes and R. Baeza-Yates, editors. *Information Retrieval: data structures and algorithms.* Prentice Hall, Englewood Cliffs, New Jersey, USA, 1992.

[G̈88] P. Gärdenfors. *Knowledge in flux: modelling the dynamics of epistemic states.* The MIT Press, Cambridge, Massachusetts, USA, 1988.

[HC68] G.E. Hughes and M.K. Cresswell. *An Introduction to Modal Logic.* Muthuen and Co. Ltd, London, UK, 1968.

[Hil65] D.J. Hillman. Topology and document retrieval operations. Studies of theories and models of information storage and retrieval, Lehigh University, July 1965.

[Kri71] S.A. Kripke. Semantical considerations on modal logic. In L. Linsky, editor, *Reference and modality*, chapter 5, pages 63–73. Oxford University Press, Oxford, UK, 1971.

[Lew81] D. Lewis. Probability of conditionals and conditionals probabilities. In W.L. Harper, R. Stalnaker, and G. Pearce, editors, *Ifs*, The University of Western Ontario Series in Philosophy of Science, pages 129–147. D.Reidel Publishing Company, Dordrecht, Holland, 1981.

[Nie88] J.Y. Nie. An outline of a general model for information retrieval. In *Proceedings of ACM SIGIR*, pages 495–506, Grenoble, France, June 1988.

[Nie89] J.Y. Nie. An Information Retrieval model based on Modal Logic. *Information Processing & Management*, 25(5):477–491, 1989.

[Por80] M.F. Porter. An algorithm for suffix stripping. *Program*, 14(3):130–137, 1980.

[RS76] S.E. Robertson and K. Sparck Jones. Relevance weighting of search terms. *Journal of the American Society for Information Science*, 27:129–146, May 1976.

[Sal68] G. Salton. *Automatic information organization and retrieval.* Mc Graw Hill, New York, 1968.

[Sta81] R. Stalnaker. Probability and conditionals. In W.L. Harper, R. Stalnaker, and G. Pearce, editors, *Ifs*, The University of Western Ontario Series in Philosophy of Science, pages 107–128. D.Riedel Publishing Company, Dordrecht, Holland, 1981.

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

[vR79] C.J. van Rijsbergen. *Information Retrieval.* Butterworths, London, second edition, 1979.

[vR86] C.J. van Rijsbergen. A non-classical logic for Information Retrieval. *The Computer Journal*, 29(6):481–485, 1986.

[vR89] C.J. van Rijsbergen. Toward a new information logic. In *Proceedings of ACM SIGIR*, Cambridge, USA, June 1989.