The troubles with using a logical model of IR on a large collection of documents

Fabio Crestani
Dipartimento di Elettronica ed Informatica
Universitá a degli Studi di Padova
Padova - Italy

Ian Ruthven
Mark Sanderson
Keith van Rijsbergen
Computing Science Department
University of Glasgow
Glasgow - Scotland

Abstract
This is a paper of two halves. First, a description of a logical model of IR known as imaging will be presented. Unfortunately due to constraints of time and computing resource this model was not implemented in time for this round of TREC. Therefore this paper's second half describes the more conventional IR model and system used to generate the Glasgow IR result set (glair1).

1 Introduction
In Information Retrieval (IR) there is no lack of models: the Boolean, the vector space, the probabilistic, and the fuzzy model are all well known. However the limitations of these models for IR cause researchers to propose new models every so often.

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 *[Salton 68]*, with a few notable exceptions *[Hillman 65]*. 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 describing 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[1] 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. Later, it was realized that it was necessary to take into consideration the uncertainty inherent in this implication. A collection of documents cannot be considered as a consistent set of statements. In fact documents in a collection could 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 $P(q|d)$. The limitation of this approach was noted by Lewis *[Lewis 81]* who showed that in *general* the probability of a conditional can not be equated to a conditional probability.

In 1986 Van Rijsbergen *[van Rijsbergen 86]* proposed the use of a non-classical conditional logic for IR. This would enable the evaluation of $P(d \rightarrow q)$ using the following logical uncertainty principle:

> "Given any two 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 $y \rightarrow x$."

The proposal initiated a new line of research (see for example *[Nie 88, Nie 89, Chiaramella 92, Bruza 93]*) but in that paper nothing was said about how "uncertainty" and "minimal" might be quantified and the suggested information-

1 In saying 'relevant to the query', what is actually meant is 'relevant to a user whose information need in expressed by a query'.

theoretic approach did not go much further than a suggestion.

A few years later, moving into Modal Logic, Van Rijsbergen proposed to estimate the probability of the conditional by a process called Imaging *[van Rijsbergen 89]*. This paper describes that proposal in more detail.

In *[Crestani 95a]* Crestani and Van Rijsbergen propose a technique called Retrieval by Logical Imaging (RbLI), that is based on the ideas suggested by Van Rijsbergen. It enables the evaluation of $P(d \rightarrow q)$ and $P(q \rightarrow d)$ by Imaging according to a possible worlds semantics where a term is considered as a possible world. This technique exploits term-term relationships in retrieval by means of an accessibility relation between worlds based on the Expected Mutual Information Measure (EMIM).

As was already indicated in the abstract however, we were unable to implement RbLI on the TREC B collection in time for the TREC-4 result submission deadline. The results submitted by us (labelled glair1) were generated from a 'traditional' ranked retrieval set up.

The structure of this paper is as follows. In the first section we present the imaging model, followed by the experimental settings for the implementation of this model. As this was not implemented in time, section 4 describes the system used to generate the substitute result set we submitted instead. The results of this set are briefly discussed in section 5 before finally concluding in section 6.

2 Retrieval by logical imaging

Logical Imaging (LI) is a technique developed in the framework of Conditional Logics that enables the evaluation of a conditional sentence without explicitly defining the operator "$\rightarrow$" (see *[Lewis 81, Stalnaker 81]*). LI is based on the Possible World Semantics *[Kripke 71]*, 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" (*[Hughes 68]*, 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 *[Lewis 81]*. 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 into a detailed explanation of the Imaging process. The interested reader can look at papers by Stalnaker *[Stalnaker 81]*, Lewis *[Lewis 81]*, and Gärdenfors *[Gardenfors 88]* 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 RbLI a model that produces a ranking of every document d in the collection based on an estimate of $P(d \rightarrow q)$. A detailed explanation of this model can be found in *[Crestani 95]*. Briefly, in RbLI we derive P_d by imaging on d assuming all the index terms t in T as possible worlds. More formally:

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

where P(t) is the "prior" 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 $I(t_d, q)$ and I(t, q) are defined as:

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

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 the following section.

In the context of the Logical Uncertainty Principle, it should be noticed that Imaging 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 occurring 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 *[Cross 94]*.

t	P(t)	I(t, d)	t_d	$P_d(t)$	I(t, q)	$P_d(t) \bullet I(t,q)$
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→q) by imaging on d

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_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 shows the evaluation of P(d→q) by imaging on d. The evaluation process is as follows:

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 $P_d \bullet I(t, q)$ for all terms (seventh column) and evaluate $P_d(q)$ by summation (bottom of seventh column).

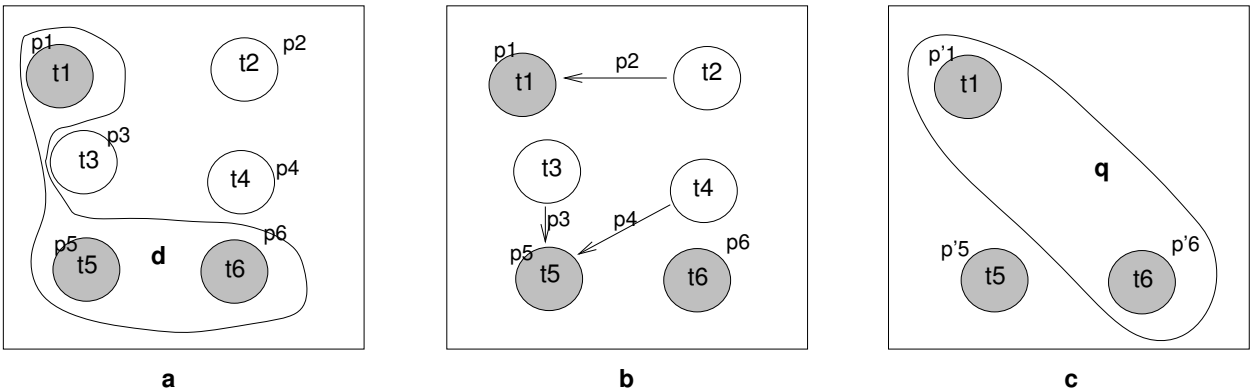

Figure 1: Graphical interpretation of the evaluation of P(d→q) by imaging on d.

A graphical interpretation of this process is depicted in Figure 1. As we said, we assume we have a measure of similarity on the term space, using it we can transfer probability from each term not occurring in the document d to its most similar one occurring in d. After the transfer of probability, terms with null probability disappear and those occurring in the query q are taken into consideration, so that their "posterior" probabilities $P_d(t)$ can be summed to evaluate $P_d(q)$.

3 Implementing retrieval by logical imaging

In order to implement the RbLI model we require:

1. A "prior" probability distribution over the index term space that should reflect the importance of each index term in the term space;

2. A measure of similarity (or alternatively a distance) between index terms;

These two requirements reflects the use of a Possible World Semantics, since they correspond to the probability distribution, and to the accessibility relation measure among the possible worlds required by Lewis in his formulation of Logical Imaging *[Lewis 81]*.

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 in IR and many models have been proposed for this purpose. The problem could be translated into finding a measure of the importance of the term in the term space, where this importance is related to the ability of the term to discriminate between relevant and not relevant documents. In IR several discrimination measures have been proposed, see for example *[van Rijsbergen 79, Robertson 76]*, and it is not clear which one should be preferred to the others. For the experiments reported in this paper we used the Inverse Document Frequency (idf), a measure often used in IR systems. It assigns a high weight and therefore a high discrimination power to terms with low and medium frequency of occurrence in the entire collection. The idf measure is defined as follows:

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

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 $\sum_t \mathrm{idf}(t) \neq 1$, however we assume it to be monotone to P(t). We can use idf instead of a proper probability function because we are only interested in a ranking of the documents of the collection, we are not interested in the exact probability values. We also chose this measure because it does not require relevance information. A discrimination measure based on some information about the relevance of a term (or document) would be much more precise, however at this stage of our work we prefer not to require relevance information. In the future we will try to use relevance information coming from such techniques as Relevance Feedback.

The problem of measuring the similarity between index terms and its use to define the accessibility among worlds is more difficult. It is important to chose the appropriate measure since much of the power of RbLI depends on it. For the experiments reported in this paper we decided to use 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, as it is 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 *[van Rijsbergen77]*. This technique makes use of co-occurrence data that can be simply derived by a statistical analysis of the term occurrences in the collection. Using this measure we can then evaluate for each term a ranking of all the other terms according to their decreasing level of similarity with it. This information is then used to determine the probability transfers for the terms in each document in the collection under consideration.

4 Implementation details

The implementations of imaging reported in Crestani and van Rijsbergen *[Crestani 95]* were performed on small collections such as Cranfield and CACM. Because of their size, it was possible to compute in a reasonable time the probability transfer of all terms for 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 our computing resources. So methods of optimising the probability transfer are being investigated. These methods include restricting the number of transfers to be performed and reducing the area within a document in which term-term similarities are calculated. Unfortunately, these optimisations were insufficient to allow completion of these experiments in time.

Rather than not submit results to TREC, it was decided some sort of result set should be sent. So a retrieval run was performed on an IR system running a 'text book' strategy, which is now described. Terms in the collection and queries first had their case normalised, then, any of these terms appearing in a stop list (taken from van Rijsbergen *[van Rijsbergen 79]*) were removed. The remaining terms were suffix stripped using the Porter stemmer *[Porter 80]*. Document terms were weighted using a tf•idf weighting scheme. The formula for idf has already been defined in a previous section, tf is defined as follows.

$$tf_{ij} = \frac{\log(freq_{ij} + 1)}{\log(length_j)}$$

$freq_{ij}$ = *the frequency of term i in document j*
$length_j$ = *the number of unique terms in document j*

The formula to calculate tf is taken from Frakes *[Frakes 92]*, tf•idf is simply, the product of the two components. Retrieval was merely a process of ranking documents in the collection based on a score. The score for each document was calculated by summing the tf•idf weights of any query terms found in that document.

5 Analysis of experimental results

Our interest in the retrieval performance of the Glasgow IR system is not particularly great as the retrieval techniques employed by this system are well known and not new. When it became clear that the imaging experiments would not be ready in time for the TREC deadline, the glair1 result set was quickly generated and submitted instead. There was no research interest in sending these substitute results, they were merely sent to provide one more set of potentially relevant documents for TREC to analyse.

In submitting a set of results from a conventional ranked retrieval set up, there was a belief (or fear?) amongst the Glasgow IR group that the glair1 result set would be one of the poorest results in TREC 4 and would provide a baseline for others. It was therefore surprising to us when told that these results were above the median of all submitted results. As no analysis of the glair1 results has been performed, one can only speculate on the factors influencing this reasonable performance. (One is tempted to say that it is slightly depressing to see this form of simple retrieval still performing well against other techniques even though it was first thought of 15 or 20 years ago.)

One of the more likely factors is the length of the queries in TREC-4. More involved strategies that have been successful in previous TREC rounds (where the queries were significantly larger) might not work so well on queries of this size, one such example might be automatic query expansion.

6 Conclusions and further work

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. Our imaging experiments are continuing and we anticipate having preliminary results soon. Trying to implement imaging on the TREC collection is proving to be compromise between the theoretical purity demanded by imaging, and the implementation problems posed by a collection of the size of TREC. We have found this compromise to be a driving force in revealing other areas of work to be investigated. Therefore despite, our failure to implement logical imaging the TREC B collection this year, we regard our participation in TREC as having been beneficial.

7 References

[Bruza 93]
P.D. Bruza, "Stratified Information Disclosure: a synthesis between Hypermedia and Information Retrieval"
Phd thesis, Katholieke Universiteit Nijmegen, The Netherlands, 1993.

[Chiaramella 92]
Y. Chiaramella and J.P. Chevallet, "About retrieval models and logic"
The Computer Journal, 35(3):233-242, 1992.

[Crestani 95]
F. Crestani and C.J. van Rijsbergen, "Information Retrieval by Logical Imaging"
Journal of Documentation, 51(1):1-15, 1995.

[Cross 94]
C.B. Cross, "Eliminative bayesianism and probability revision"
Unpublished paper, August 1994.

[Frakes 92]
W.B. Frakes & R. Baeza-Yates, "Information Retrieval: data structures and algorithms"
Prentice-Hall

[Gärdenfors 88]
P. Gärdenfors, "Knowledge in flux: modelling the dynamics of epistemic states'
The MIT Press, Cambridge, Massachusetts, USA, 1988.

[Hillman 65]
D.J. Hillman, "Topology and document retrieval operations"
Studies of theories and models of information storage and retrieval, Lehigh University, July 1965.

[Hughes 68]
G.E. Hughes and M.K. Cresswell, "An Introduction to Modal Logic"
Muthuen and Co. Ltd, London, UK, 1968.

[Kripke 71]
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.

[Lewis 81]
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.

[Nie 88]
J.Y. Nie, "An outline of a general model for information retrieval"
In Proceedings of ACM SIGIR, pages 495-506, Grenoble, France, June 1988.

[Nie 89]
J.Y. Nie, "An Information Retrieval model based on Modal Logic"
Information Processing & Management, 25(5):477-491, 1989.

[Porter 80]
M.F. Porter, "An algorithm for suffix stripping"
Program, Vol. 14, Num. 3, Pages 130-137

[van Rijsbergen 79]
C.J. van Rijsbergen, "Information Retrieval"
Butterworths, London, second edition, 1979.

[van Rijsbergen 86]
C.J. van Rijsbergen, “A non-classical logic for Information Retrieval”
The Computer Journal, 29(6):481-485, 1986.

[van Rijsbergen 89]
C.J. van Rijsbergen, “Toward a new information logic”
In Proceedings of ACM SIGIR, Cambridge, USA, June 1989.

[Robertson 76]
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.

[Salton 68]
G. Salton, “Automatic information organization and retrieval”
Mc Graw Hill, New York, 1968.

[Stalnaker 81]
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.