A Study of Probability Kinematics in Information Retrieval

F. Crestani*
C.J. van Rijsbergen
Department of Computing Science
University of Glasgow
Glasgow G12 8QQ, Scotland

July 9, 1997

Abstract

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

Categories and Subject Descriptors: F.1.1 [Computation by Abstract Devices]: Models of Computation; H.3.3 [Information Storage and Retrieval]: Information Search and Retrieval;

General Terms: Probabilistic Modelling, Logics.

*Previously at Dipartimento di Elettronica e Informatica, Universita' di Padova, Italy.

1 Introduction

In Information Retrieval (IR), probabilistic modelling means the use of a model that ranks documents in decreasing order of their evaluated or, more often, estimated probability of relevance to a user's information need expressed by a query. In an IR system based on a probabilistic model, the user is always guided to examine first the documents that are the most likely to be relevant to her need, and if the system worked well these documents should be at the top of the ranked retrieved list.

The use of a probabilistic model in IR assures that we can obtain "optimal retrieval performance" once we rank documents according to their probability of relevance with regard to a query [Robertson, 1977]. However, this principle, called *The Probability Ranking Principle*, refers only to "optimal retrieval", which is different from "perfect retrieval". Optimal retrieval can be defined precisely for probabilistic IR because optimality can be proved theoretically, owing to a provable relationship between ranking and the probabilistic interpretation of precision and recall [Robertson, 1976]. Perfect retrieval relates to the objects of the IR systems themselves, i.e. documents and information needs, but as IR systems use representations of these objects, perfect retrieval is not an appropriate goal for computer-based systems and cannot be achieved experimentally. Despite that and despite a few criticisms [Cooper, 1994], probabilistic models based on the Probability Ranking Principle have been shown to give the highest levels of retrieval effectiveness currently available [Fuhr, 1992].

Although there are some operative IR systems based on probabilistic or semi-probabilistic models, there are still obstacles to getting probabilistic models accepted in the commercial IR world. One major obstacle is that of finding methods for estimating the probabilities of relevance that are both effective and computationally efficient. Past and present research has made much use of formal probability theory and statistics in order to solve the problems of estimation, see for example [Croft and Harper, 1979, Fuhr and Buckley, 1991, Wong and Yao, 1989]. In mathematical terms the problem consists of estimating the probability $P(R \mid q, d)$, that is the probability of relevance given a query q and a document d. This estimate should be performed for every document in the collection, and documents should then be ranked according to this measure. This is a difficult task because of the large number of variables involved in the representation of documents in comparison with the small amount of feedback data available about the relevance of documents, a problem sometimes referred to as the "curse of dimensionality" [Efthimiadis, 1996, van Rijsbergen, 1979].

In 1986 Van Rijsbergen [van Rijsbergen, 1986] proposed to use an estimation techniques based on the use of non-classical conditional logic. This enables the

estimation of $P(R \mid q, d)$ by the evaluation of $P(d \rightarrow q)$, therefore using the probability of a conditional to estimate the conditional probability.

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

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

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

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

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

According to Van Rijsbergen's view, the logical implication $d \rightarrow q$ is not one of material implication, the usual truth-functional connective $d \supset q$, which is always true in all cases except when d is true and q is false, but is based on a non-classical notion (see Section 3.3). The evaluation of the probability of the implication should be based on the following *logical uncertainty principle*:

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

The logical uncertainty principle initiated a new line followed by many researchers; see for example the work of Nie [Nie, 1989], Chiaramella and Chevallet [Chiaramella and Chevallet, 1992] Bruza [Bruza, 1993], and Huibers [Huibers *et al.*, 1996]. However, in the original 1986 paper Van Rijsbergen did not provide an indication about how "uncertainty" and "minimal" might be quantified. Only a few years later Van Rijsbergen proposed to estimate the probability of the conditional by a process called *Logical*

Imaging (or simply Imaging), but without explicitly defining a technique that could be used operatively [van Rijsbergen, 1989]. The Imaging technique was explored by Crestani and Van Rijsbergen and a model called *Retrieval by Logical Imaging* has now been defined in detail [Crestani and van Rijsbergen, 1995].

In this paper we explore further the use of the probability of a conditional, namely $P(d \rightarrow q)$, to estimate the conditional probability $P(R \mid q, d)$. We propose the use of a model called *Retrieval by General Logical Imaging*, based on a generalisation of the Retrieval by Logical Imaging model. We analyse and compare the probability kinematics of imaging and general imaging with that of two more classical probabilistic models: the Retrieval by Joint Probability model and the Retrieval by Conditional Probability model. These two models are at the basis of many IR probabilistic models currently in use.

The paper is structured as follows. Section 2 describes the representation model for documents and queries that we will use in the rest of the paper. Section 3 describes the probability kinematics of four different retrieval models and explains in detail Retrieval by Logical Imaging and Retrieval by General Logical Imaging. The retrieval performance of these four models is then studied using the probability and similarity functions described in Section 6 and the experimental settings described in Section 5. The results are presented, compared, and discussed in Section 7. Section 8 reports the conclusions of the experimental investigation and the pitfalls in the proposed models.

2 The representation space

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

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

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

The problem of *estimating* $P(R \mid q, d)$ is tackled in this paper from a theoretical point of view. In the past researchers have tried to estimate $P(R \mid q, d)$ in many different ways. Often, researchers have had to resort to *ad hoc* estimation techniques, very much dependent on experimental tuning of the parameters of their models. In Section 3 we will report on four theoretical retrieval models. Without entering into a discussion on the process of parameter estimation, we will analyse their differences and will draw some interesting conclusions that could be useful in directing new research in this area.

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

2.1 Possible World Semantics and Logical Imaging

Possible World Semantics was introduced by Kripke [Kripke, 1971] in the context of Modal Logic. In this semantics the truth value of a logical sentence is evaluated in the context of a "*world*". The word "world" has been used like this by a number of logicians and seems to be the most convenient one, but perhaps some such phrase as "conceivable or envisageable state of affairs" (used in [Hughes and Cresswell, 1968], p. 75) would convey the idea more clearly. Possible World Semantics has been used in Modal systems to give a semantics for Necessity (where a sentence is true in every possible world) and Possibility (where a sentence is true in at least one possible world)[3].

Without entering into the details of this semantics, one of the main advantages of Possible World Semantics is that it enables the evaluation of the truth value of a conditional sentence without explicitly defining the operator "$\rightarrow$" [Lewis, 1986]. What it requires is a clustering on the space of events (worlds) by means of a primitive relation of neighbourhood. The clustering then enables us to define an *accessibility relation* that is necessary for the evaluation of the conditional sentence. According to the Possible World Semantics the truth value of the

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

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

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 [Stalnaker, 1981]. Ties at this stage, if they occur, are broken at random, to ensure the uniqueness of the closest world (but see further on for a generalisation). The passage from one world to another world can be regarded as a form of belief revision, and the passage from a world to its nearest neighbour is equivalent to the least drastic revision of one's beliefs. Using this process is a mean of implementing the logical uncertainty principle described in Section 1.

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

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

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

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

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

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

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

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

This is the technique called *Logical Imaging* and was first proposed by Stalnaker [Stalnaker, 1981]. The arguments related to the existence of the most similar world to w are addressed in [Lewis, 1986, Gärdenfors, 1988]. We will not raise them here since we will be imposing a metric on W that will enable us always to find at least one most similar world to any given world w.

Imaging has been extended by Lewis [Lewis, 1981] to the case where there is a probability distribution on the worlds. Let us assume it follows the classical rules of probability, and in particular:

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

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

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

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

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

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

where:

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

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

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

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

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

where $P_y(x)$ is the new probability distribution, called "posterior" probability, derived from P by imaging on y. In other words, the probability of the conditional is the probability of the consequent after imaging on the antecedent. The proof is reported in [Lewis, 1981]. The interested reader can also look at the following papers by Stalnaker [Stalnaker, 1981], Gärdenfors [Gärdenfors, 1982] and Cross [Cross, 1994] for more details of the imaging process.

In 1988 Gärdenfors proposed a generalisation of the imaging process [Gärdenfors, 1988]. The generalisation originated from an attempt to overcome one of the restrictive assumptions Lewis made for Stalnaker's semantics of conditionals [Lewis, 1981]. The assumption is related to the "uniqueness" of the world w_y, that is the uniqueness of the most similar y-world to w. The generalisation that Gärdenfors proposed does not rely on this assumption[4]. The starting point is the use of a probability function $P^w(w')$ to represent the belief in the world w' given that the world w is certain. This probability function enables us to evaluate the (degenerated) probability function $P^w(y)$ that can be used to represent the fact that in any possible world w a sentence y can be true to a certain extent. The probability function $P^w(y)$ is derived from a probability distribution $P^w(w')$ over the possible worlds in such a way that:

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

Lewis called the probability function $P^w(y)$ "opinionated" because "it would represent the beliefs of someone who was absolutely certain that the world w was actual and who therefore held a firm opinion about every question" (see [Lewis, 1981], p. 145). Gärdenfors generalised imaging by considering the fact that, instead of having $P^w(y) = 1$ only for a single world w_y, we can have $0 \leq P^w(y) \leq 1$ for a set of worlds:

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

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

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

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

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

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

where:

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

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

It could be proved, with a demonstration similar to the one reported in [Lewis, 1981], p. 142, that this new probability distribution is the posterior probability distribution derived from the prior probability P by *General Logical Imaging* on y. In other words, this new probability distribution can be obtained by transferring the probability from every world w to the worlds in W_y, the set of most similar (closest) worlds to w where y is true. The transfer of probability is performed according to the opinionated probability function $P^w(y)$. It is easy to prove that Lewis' imaging is just a special case of general imaging when $P^w(y) = 1$ for just one w.

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

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

Again, it can be demonstrated that:

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

The evaluation of $P_y(x)$ either by imaging or general imaging causes a shift of the original probability P from "not-y-worlds" to "y-worlds" to derive a new probability distribution P_y. Since the transfer of probabilities is directed towards the closest y-worlds, this technique is just what it is needed to implement the logical uncertainty principle described in Section 1. The probability revision is in fact minimal with regards to the accessibility relation, that is to say, it minimises the total amount of the distance covered in the probability transfer. In Sections 3.3 and 3.4 we will explain how we can use this result in the context of IR, but first let us examine how we can use the Possible World Semantics to model the term space.

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

Figure 1: The classical geometrical space semantics for the term space

2.2 The term space

One of the best known IR models is the Vector Space Model (VSM) [Salton, 1968]. In VSM a document is represented by means of a vector whose elements represent the presence/absence of certain features in the document representation, for example, the presence or absence of index terms. Considering the binary case, for simplicity of exposition, a 1 in a particular position of the vector indicates the presence in the document representation of the feature associated with that position , while a 0 indicates its absence. The document representation space is therefore multidimensional, with as many dimensions as the number of features used for representing documents. Documents and queries are represented in this space as vectors. The semantics of the VSM is therefore that of a multidimensional geometrical space. The topology and the metrics of this space enable the evaluation of the RSV of a document with regards to a query as a distance measure. Many IR models use a similar representation space.

Here we use the same representation space but a different semantics. The semantics of our representation space is based on the Possible World Semantics. We use the Possible World Semantics in the context of IR by considering a term as a possible world, a view that was proposed in [Crestani and van Rijsbergen, 1995]. According to this view, a term is represented as a "vector of documents". Intuitively this can be understood as "if you want to know the meaning of a term then look at all the documents in which that term occurs". This idea is not new in IR (see for example [Qiu and Frei, 1993]) and it has been widely used for the evaluation of term–term similarity (see Section 6).

More formally, let us assume our representation space is made of a set of index terms T, we will call it a *Term Space*. The set of index terms T is our set of possible worlds. We also assume we have a document collection D where each document d is represented using terms in T, as depicted in the representation matrix in figure 1. According to our semantics, in order to determine if a document is true or not in the context of a term it is sufficient to transpose the

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

Figure 2: Application of the Possible World Semantics to the term space

representation matrix and consider a document true in the context of a term if it uses that term in its representation. The matrix depicted in figure 2 can therefore be interpreted as representing the truth values of documents in the context of terms.

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

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

3 Probability kinematics in IR

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

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

3.1 Retrieval by Joint Probability

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

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

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

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

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

RbJP is the simplest approximation to $P(R \mid q,d)$, but it is used by many IR models. In fact it is the archetype of many IR models currently in use. Most IR models that are based on the evaluation of similarity between documents and query are based upon the idea of a joint probability measure. Both Dice's and Jaccard's coefficients (see [van Rijsbergen, 1979], p. 39) are based on it, as can be seen once we remove the normalisation factors.

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

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

where $\parallel S \parallel$ indicates the cardinality of the set S.

The results of the last formula is monotone to the *Retrieval by Simple Matching* value and with $k = 1$ we obtain the "coordination level coefficient" (the number of terms the query has in common with the document, see [van Rijsbergen, 1979], p. 97) one of the oldest IR models. From the RbJP formula we can also obtain Dice's and Jaccard's coefficients just by assigning to k different normalisation factors. The *Cosine Correlation* used by the VSM is also a normalised version of RbJP, it is in fact the same formula with a Euclidean norm:

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

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

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

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

Table 1: Example of the evaluation of $P(q,d)$

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

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

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

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

This corresponds to the *Cosine Correlation* using the "$tf \cdot idf$" weighting scheme as defined in [Salton and Yang, 1973]. We should notice that $P(q,d)$ is no more a measure of probability, since it does not give values between 0 and 1, however this problem could be simply solved introducing a normalisation factor.

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

To show how we evaluate the RSV in the case of RbJP, we refer to Table 1, where we report the evaluation of $P(q,d)$ for a particular document and query. The evaluation process is the following:

1. Identify the terms occurring in the document d.
2. Identify the terms occurring in the query q.

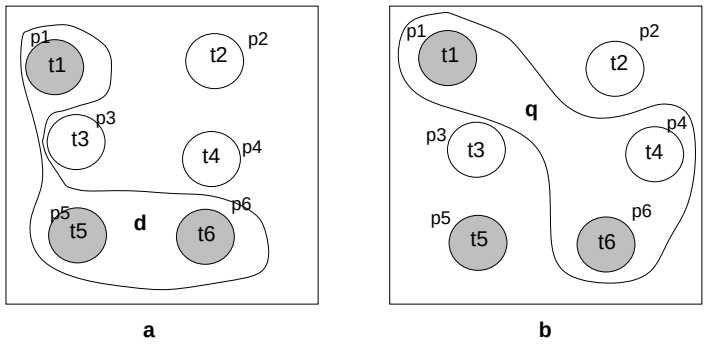

Figure 3: Graphical interpretation of the evaluation of $P(q, d)$.

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

A graphical interpretation of RbJP using the Possible World Semantics is given in figure 3, where each term is represented by a world with its prior probability measure expressing the importance of the term in the term space T. The shaded terms are those occurring in document d (figure 3(a)). The value of $P(q, d)$ is obtained by summing the probability of all terms occurring both in the document and in the query representations, that is summing the probabilities of the shaded terms also occurring in q (figure 3(b)).

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

3.2 Retrieval by Conditional Probability

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

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

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

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

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

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

$$\lambda_d = \frac{\sum_{t \notin d} P(t)}{\sum_{t \in d} P(t)}$$

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

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

The normalisation enables the prior probability to be revised in accordance with the characteristics of the particular document under consideration. Thus, the transfer of probabilities that takes place in RbCP provides the minimal revision of the prior probability necessary to make d certain without distorting the profile of probability ratios. In fact, the posterior probability is proportional to the prior probability, so leaving constant the ratio of probabilities associated with the terms after the contraction of the term space due to the fact that the conditional event d has become certain.

Wong and Yao in [Wong and Yao, 1995] demonstrated that most probabilistic models of IR can be explained using $P(q \mid d)$. The difference between the various models is given by the different ways $P(q \mid d)$ can be evaluated.

Table 2 reports an evaluation of $P(q \mid d)$. The evaluation process is the following:

1. Identify the terms occurring in the document d.
2. Evaluate the posterior probability $P_d(t)$ by transferring the probabilities from terms not occurring in the document to terms occurring in it. The probabilities are transferred in a proportional way, so that each term occurring in the document d receives a portion of the sum of the probability of the terms not occurring in the document proportional to its prior probability.

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

Table 2: Example of the evaluation of $P(q \mid d)$

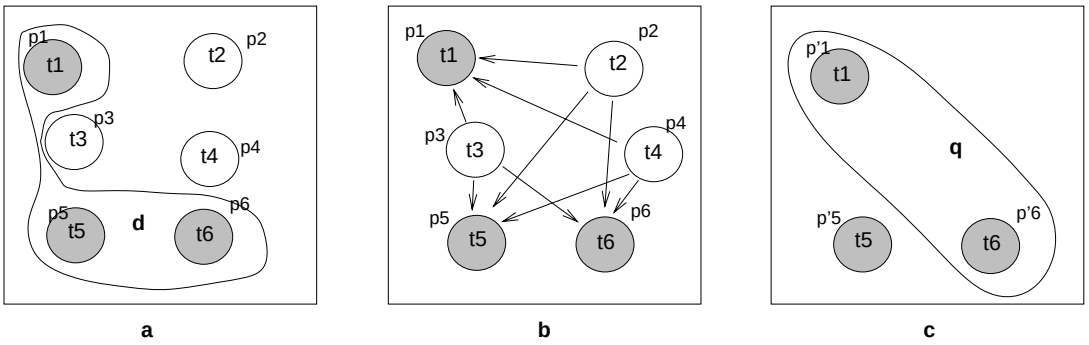

Figure 4: Graphical interpretation of the evaluation of $P(q \mid d)$.

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

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

It is interesting to see a graphical interpretation of probability kinematics induced by RbCP. In figure 4(a) each term is represented by a world with its prior probability. The shaded terms occur in document d. The conditioning process transfers the probability from terms not occurring in d to those occurring in it as depicted in figure 4(b). In figure 4(c) the terms with null probability disappear and only those terms occurring in the query q are taken into consideration to evaluate $P(q \mid d)$.

3.3 Retrieval by Logical Imaging

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

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

The motivation behind this approach is related to the underlying definition of Relevance (R). We accept a logical notion of relevance, in accordance with the work of Cooper [Cooper, 1971] and Van Rijsbergen [van Rijsbergen, 1986]. Relevance is defined as the truth value of the implication $d \rightarrow q$. An equivalent interpretation of the truth of $d \rightarrow q$ is to consider d and q as events, then the "satisfaction" of a document described by d entails the satisfaction of a query described by q. The satisfaction of a document d means that the logical expression d is true in the current retrieval situation. A particular case in which d is true is when a document corresponding to d is retrieved, so a slightly narrower interpretation of the truth of $d \rightarrow q$ is: "the retrieval of d leads to the satisfaction of q". It is known in IR that relevance is often uncertain due to the uncertainty in the description of the contents of documents and queries, therefore we cannot talk about the truth of $d \rightarrow q$, but of the degree of certainty (or uncertainty) of the truth value. This leads us to talk about $P(d \rightarrow q)$.

Wong and Yao [Wong and Yao, 1995] suggested estimating $P(d \rightarrow q)$ by $P(q \mid d)$. The limitations of this approach are known in the area of logics by the name of "triviality results", and were well illustrated by Lewis in [Lewis, 1981]. According to these results $P(d \rightarrow q)$ and $P(q \mid d)$ would be equal only in certain extreme cases that are so simple that they can be considered "trivial"[5]. These results excluded that conditional probabilities could be used as a probabilistic logic dealing with conditionals. As a consequence, Lewis suggested estimating the probability of a conditional using Logical Imaging.

Retrieval by Logical Imaging (RbLI) is the model that estimates $P(R \mid q, d)$ by $P(d \rightarrow q)$, where the latter is evaluated using logical imaging. A detailed explanation of the RbLI model can be found in [Crestani and van Rijsbergen, 1995].

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

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

where t_d is the closest term to t for which d is true, or in other words, the most similar term to t that also occurs in the document d. The application

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

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

Table 3: Example of the evaluation of $P(d \rightarrow q)$ by imaging on d

of imaging to IR requires an appropriate measure of similarity over the term space to enable the identification of t_d. This is the equivalent of the accessibility relation described by Lewis in [Lewis, 1981]. The measure of similarity used in the evaluations reported in this paper is described in Section 6.

RbLI implements Van Rijsbergen's Logical Uncertainty Principle because it provides the minimal revision of the prior probability in the sense that it involves no gratuitous movement of probability from one world to dissimilar worlds. The revision of the prior probability necessary to make d certain is obtained by adopting the least drastic change in the probability space. This is achieved by transferring probabilities from each term not occurring in the document d to its closest (most similar) term occurring in d, so that the total amount of the distance covered in the transfer is minimal. A detailed comparison between other forms of conditionalisation and imaging can be found in [Cross, 1994].

For a practical example of the evaluation of RbLI let us suppose we have the same query q and document d of the previous sections. Table 3 reports the evaluation of $P(d \rightarrow q)$ by imaging on d. The evaluation process is the following:

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

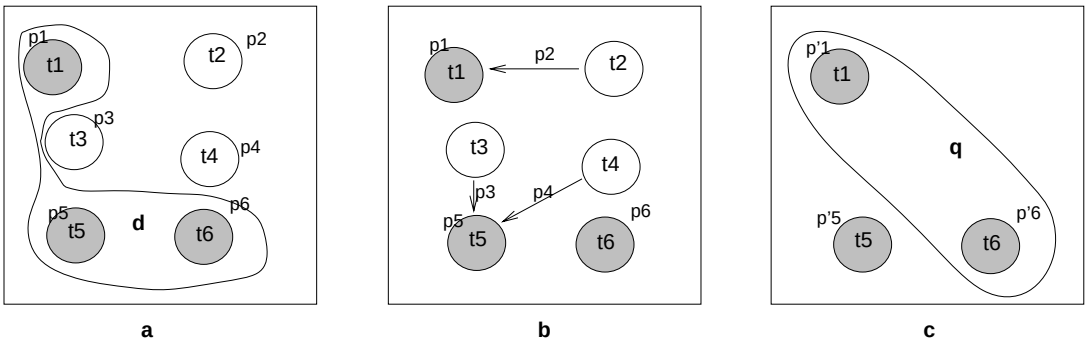

Figure 5: Graphical interpretation of the evaluation of $P(d \to q)$ by imaging on d.

A graphical interpretation of this process is depicted in figure 5. 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 (figure 5(b)). After the transfer of probabilities, terms with null probability disappear and those occurring in the query q are taken into consideration, so that their posterior probabilities $P_d(t)$ are added together to evaluate $P_d(q)$.

3.4 Retrieval by General Logical Imaging

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

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

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

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

The opinionated probability function $P_d^{t'}(t)$, defined in Section 2.1, determines the amount of probability to be moved from t' to the term t belonging to the set T_d, where $T_d \subset T$ is the set of all terms occurring in document d. Such a function depends on the particular document on which general imaging is performed and

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

Table 4: Example of the evaluation of $P(d \to q)$ by general imaging on d

on the particular term from which we want to transfer the probability. The number of opinionated probability functions required is then equal to the product of the number of documents multiplied by the number of terms. This number could be very high. We are currently working on this problem and we plan to use contextual information together with similarity information to determine it, with the document giving the context in which the similarity is measured. However, for the tests reported in this paper we will make some strong assumptions:

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

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

These two assumptions enable us to use a single opinionated probability function for every term in the term space. We plan in the near future to remove first the second assumption and perform a transfer of probability that is related to the value of similarity between two terms, and later remove also the first assumption to make this transfer dependent on the context set by the document. In the tests reported in this paper, however, we will make use of a very simplistic opinionated probability function. Such a function is described in detail in Section 6.

Table 4 reports an example of the evaluation of $P(d \to q)$ by general imaging on d. The evaluation process is the following:

1. Identify the terms occurring in the document d.

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

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

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

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

A graphical interpretation of this process is shown in figure 6. As can be seen in the picture, in the case of RbGLI the transfer of probability is performed from each term not occurring in the document d to the k_t most similar terms occurring in d. In the example $k_t = 2$ for every term, but k_t can be any other integer number so that $1 \leq k_t \leq l_t$, where l_t is the number of documents in which the term t occurs. The value of k_t is in theory different for every term, but can be set to a constant k independent of the term t. This setting simplifies considerably the evaluation of RbGLI. If $k_t = 1$ for every term, then RbGLI defaults down to RbLI. If $k_t = l_t$ for every term, then the transfer looks similar to the one produced by RbCP. However, note that the probability transfer in RbGLI is performed by taking into account the similarity between terms and not the ratio of prior probabilities as in the case of RbCP. Gärdenfors demonstrated in [Gärdenfors, 1988] that it is not possible to find any prior probability distribution for which the transfer of probability induced by general imaging is equivalent to that induced by conditional probability.

4 Related work

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

In [Amati and Kerpedjiev, 1992] Amati and Kerpedjiev proposed two logical models for IR. One of them is based on conditional logic and makes use of imaging

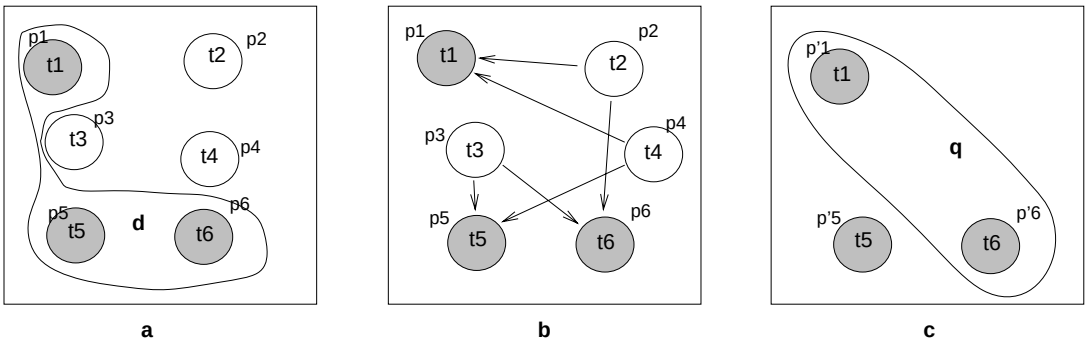

Figure 6: Graphical interpretation of the evaluation of $P(d \rightarrow q)$ by general imaging on d.

for the evaluation of $P(d \rightarrow q)$ and $P(q \rightarrow d)$. However, they proposed two different semantics for the evaluation of the two conditionals. For the evaluation of $P(d \rightarrow q)$ they consider a term as a world, while for the evaluation of $P(q \rightarrow d)$ they consider a document as a world. We see a difficulty in this latter approach because the event d in the conditional statement $q \rightarrow d$ is also interpreted as a world. To deal with this difficulty one would have to make explicit the difference between a document as a fictive object existing in its own right and a partial description of such an object (as in [Lalmas and van Rijsbergen, 1993]). Rather than do this we have adopted a different approach.

Sembok and Van Rijsbergen [Sembok and van Rijsbergen, 1993] proposed a relevance feedback technique based on the use of imaging. Again, the perspective of a document as a world is used, which gives the same problem as before. Moreover, the similarity between documents is evaluated by means of clustering using nearest neighbours. The similarity measure used for the clustering on the document space is based on Dice's coefficient, a very simple similarity measure. We think that since most of the power of imaging relies on the correct identification of the closest possible world, it is very important to use the best possible similarity measure for the job. In Section 6 we describe the measure we used for the evaluation reported in this paper.

Nie, first in [Nie, 1992] and later in [Nie *et al.*, 1995], uses imaging to include in the retrieval process such contexts as user knowledge, domain knowledge, intentions, and so on. In his model both documents and queries are sentences. Possible worlds represent different states of the data set, for example possible states of knowledge that can be held by users. A document d is true in a world w if the document is "consistent" (the term is used here in a broad sense) with the state of knowledge associated with that world. Worlds differ because they represent different states of knowledge and, given a metric on the world space, we can identify the closest world to w for which d is true. Imaging can then be

used for the evaluation of the certainty of the implication $d \rightarrow q$. Nie's approach takes a view similar to the one followed in this paper. Both approaches consider a world as an informative entity, in the context of which a document or a query need to be checked for consistency. The major advantage of Nie's model is that it enables user modelling and therefore the evaluation of a user oriented measure of relevance, while RbLI and RbGLI only takes into account a system evaluated relevance.

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

Wong and Yao [Wong and Yao, 1991, Wong and Yao, 1995] demonstrated that most of the IR models in use at present can be explained in terms of the formula $P(E \rightarrow H)$ that is evaluated as $P(H \mid E)$. The latter formula evaluates the degree of confirmation (or belief, according to the view taken) of the sentence H given evidence E. Conventional IR models can be obtained by associating either d or q to H or E, and by defining different ways of evaluating the probabilities via probabilistic inference on a concept space. Concepts are considered disjoint elements of the representation space, or are transformed in such a way to be disjoint. Terms are basic concepts.

Another important result is reported in [Wong and Yao, 1995] p. 58 where Wong and Yao show that their model, called "probabilistic inference model", subsumes the probabilistic model. Both Fuhr's probabilistic independence indexing model [Fuhr, 1989] and binary probabilistic independence retrieval model [van Rijsbergen, 1979] can be explained in terms of the probabilistic inference model. Since the probabilistic inference model is based on the concept of conditional probability, then also Fuhr's and the binary probabilistic independence models are based on the same kinematics of probabilities. The amount of probability moved from one concept to another may change, but the principle remains the same: the transfer of probabilities provides the minimal revision of the prior probability that is necessary to make the evidence E certain without distorting the profile of probability ratios on the representation space. In this paper we show that there are other types of probability kinematics that implement the Logical Uncertainty Principle in different ways. Moreover, the view taken by Wong and Yao is purely probabilistic. For them, only probabilistic inference is used for the evaluation of the uncertainty of the implication $E \rightarrow H$. We extend that view by taking into consideration a semantics of the representation space based on Possible World Semantics, which enables the evaluation of $P(E \rightarrow H)$ in a less restrictive way than does pure probabilistic inference. Thus we think that the use of the Possible World Semantics enables us to design and deal with different and more complex models of probability kinematics, like our imaging and general imaging.

We would also like to mention the work done in the context of expanding a query by adding terms that are semantically similar to those originally present. There are some similarities between our work on probability kinematics and work that has been done by others on query expansion. The transfer of probabilities that RbLI performs could be regarded as a way of expanding the terms present in the document with terms that are similar to them but not present. Work in this direction has been carried out by many researchers, for example Qiu and Frei [Qiu and Frei, 1993] and Voorhees [Voorhees, 1994]. However, the similarities between our work and query expansion are not easy to assess. There is nothing analogous to RbGLI in the context of query expansion. Moreover, once we move out from the Possible World Semantics, it is very difficult to interpret the consequences of imaging in the context of natural language, while it is easy to interpret query expansion. A study of the implications of imaging in the context of sense resolution is currently under way. Some first results have appeared in [Crestani *et al.*, 1996]. Until the natural language semantics of imaging is fully understood, it is not possible to assess clearly the differences between imaging and query expansion or other query modification techniques.

5 Experimental analysis

So far we compared four probabilistic retrieval models using a common representation space and a common semantics. The comparison was mainly on theoretical grounds and was meant to show what happens at retrieval time to probabilities assigned to the elementary objects of the representation space. We showed that these four models induce different probability kinematics in the term space. Now, an obvious question comes to mind: which of the four models is the best in practical terms? The theoretical analysis we perform allows us to see the differences between the various models, but does not allow us to say which is the best. We therefore decided to perform some performance testing to compare the retrieval of the four models. These tests are only meant to show existing significant differences in the retrieval performance of the four different models. We decided to avoid using "ad hoc" indexing and normalisation schemes or adaptations of existing IR systems, since they could throw the comparison out of "balance". We used the retrieval models as they have been described in section 3. Our results cannot therefore be compared with the results achieved by other IR systems using the same data, and they only have a comparative significance in the framework of our testing.

In order to study and compare the retrieval effectiveness of the four models under consideration we performed a series of tests using some standard test collections. We used three test collections that have been extensively studied and used in the

Data	*Cranfield*	*CACM*	*NPL*
documents	1400	3204	11429
queries	225	52	93
terms in doc.	5000	7121	7492
terms in query	274	356	337
avg. doc. length	53.61	24.26	19.96
avg. query length	8.95	11.5	7.15
avg. rel. doc.	7.28	15.31	22.41

Table 5: Test collections data

field of IR: the *Cranfield 1400*, the *CACM*, and the *NPL* test collections. The characteristics of these three test collections are described in many papers (see for example [Cleverdon *et al.*, 1966, Sparck Jones and van Rijsbergen, 1976]). A summary of the main characteristics is reported in Table 5.

The results of our test are presented using the standard evaluation technique used in IR. Precision and recall tables have been evaluated using their standard definition [van Rijsbergen, 1979]. The method of linear interpolation has been used to determine the standard values corresponding to intervals of 10% in the recall figures.

6 Prior probability, similarity and opinionated probability function

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

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

The problem of determining an appropriate prior probability distribution over the set of index terms is one of the oldest problems of IR and many ways have been proposed for this purpose. The problem could be translated into finding a

measure of the importance of a term in the term space, where this importance is related to the ability of the term to discriminate between relevant and not relevant documents. The importance of the term in the term space seems a reasonable rationale for a probability function. Several discrimination measures have been proposed, and a few examples can be found in [van Rijsbergen, 1979, Robertson and Sparck Jones, 1976]. For the tests reported in this paper we used the *Inverse Document Frequency*, a measure which assigns high discrimination power to terms with low and medium collection frequency. Strictly speaking, this is not a probability measure since $\sum_t idf(t) \neq 1$, however we can assume it to be monotone to $P(t)$. We can use this estimate because we require only a ranking of the documents in response to a query, not the exact probability values.

The problem of defining a measure of similarity between terms and the use of such a measure for defining the accessibility among worlds is more difficult, although it has been addressed by many researchers in the past, in the fields of IR [Wong *et al.*, 1993, Voorhees, 1993, Srinivadsan, 1992] and Natural Language Processing [Church and Hanks, 1989, Brown *et al.*, 1992]. It is very important to chose a good measure since much of RbLI and RbGLI depends on it. In this paper we decided to use the *Expected Mutual Information Measure* ($EMIM$), because it has been used with success in the past by one of the authors [van Rijsbergen, 1977], and because it is a well accepted measure in Lexicography [Church and Hanks, 1989].

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

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

where t_i and t_j are terms.

We can estimate $EMIM$ between two terms using the technique proposed by Van Rijsbergen in [van Rijsbergen, 1979]. This technique makes use of co-occurrence data that can be derived by a statistical analysis of the term occurrences in the collection. Using $EMIM$ we can then evaluate for every term a ranking of all the other terms according to their decreasing level of similarity with it. We store this information in a file which is used at run-time to determine for every term t its closest term occurring in d, that is t_d.

The problems related to defining an appropriate *opinionated probability function* have already been introduced in Section 3.4. This third experimental requirement is a heavy one. The problem of finding a good opinionated probability function is still open. We do not tackle it in this paper. In the tests reported in Section 7

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

Table 6: Comparison of the average precision of the four models with different test collections

we use a discrete monotonically decreasing transfer function that transfers from a term t a decreasing fraction of its probability to all the other terms in the term space once they are ordered in decreasing order of similarity. In particular, to simplify computations, in the evaluation of $P(d \rightarrow q)$ by general imaging on d, from each term not occurring in d we transfer probability only to the first 10 most similar terms occurring in d. Once terms are ordered in decreasing order of similarity with t, the probability transfer function we use works in such a way that the ith term gets double of what the $(i+1)$th gets. In the future we intend to use a more complex function that takes into account the contextual information provided by the particular document on which general imaging is performed. The opinionated probability function will be based on a term-term similarity measure evaluated in the context of that document.

7 Evaluation

We performed a comparative evaluation of the retrieval effectiveness of the four models presented above using the document collections and the experimental settings reported in Section 6 and 5.

We do not enter into the details of the evaluation since it is beyond the scope of this paper, suffice to say that the actual computations of the RSV used for obtaining the figures reported in this section are very similar to the ones reported in the examples in Section 3. The only modifications to the techniques described earlier were introduced in order to reduce the number of computations necessary at run-time for the probability transfer[6]. These modifications have already been described in detail in [Crestani and van Rijsbergen, 1995] and are not reported here for reasons of space. Here we only compare the results obtained by each model to draw some plausible conclusions.

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

Table 6 reports the average precision values obtained by the different models on different test collections. It also shows the percentage increase in the average precision gained by using the different models. It can be easily seen that the average precision increases consistently from RbJP to RbGLI in all three document collections, although the increase rate is variable.

The results displayed in the Recall/Precision graphs in figures 7, 8 and 9 show that the performance of RbGLI are slightly higher than those obtained by any other model, with RbJP at the lowest level of performance.

From the results we can observe that:

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

These findings are consistent over the three document collections.

Despite the fact that we are using a simple term weighting schema and that we are experimenting with small test collections, we think we can nonetheless conclude that the probability kinematics of IR probabilistic retrieval models is worthy of further study. An exploration of the kinematics of probabilities in IR can help discover interesting properties of existing and new retrieval models, and can provide pointers for further study on how to improve the design of new IR models. An interesting result from our study on the kinematics of the four models presented is that it is possible to obtain higher levels of retrieval effectiveness by taking into consideration the similarity between the objects involved in the transfer of probability. However, the similarity information should not be use too drastically (like in RbLI) since similarity is often based on co-occurrence and such a source of similarity information is itself uncertain. A way of partially dealing with this latter uncertainty would be to contextualise the similarity information to make it document dependent. This is a line of research we will investigate using Natural Language Processing techniques. Some initial results of this work are reported in [Crestani *et al.*, 1996].

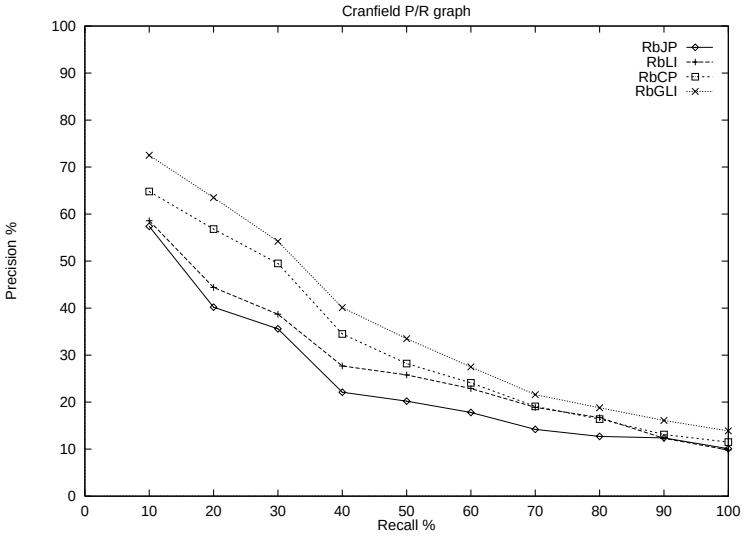

Figure 7: Precision and recall graphs for the Cranfield test collections

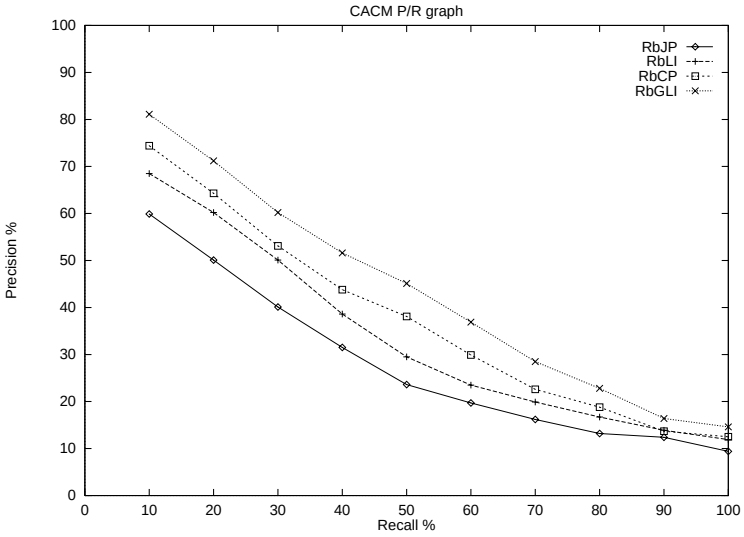

Figure 8: Precision and recall graph for the CACM test collection

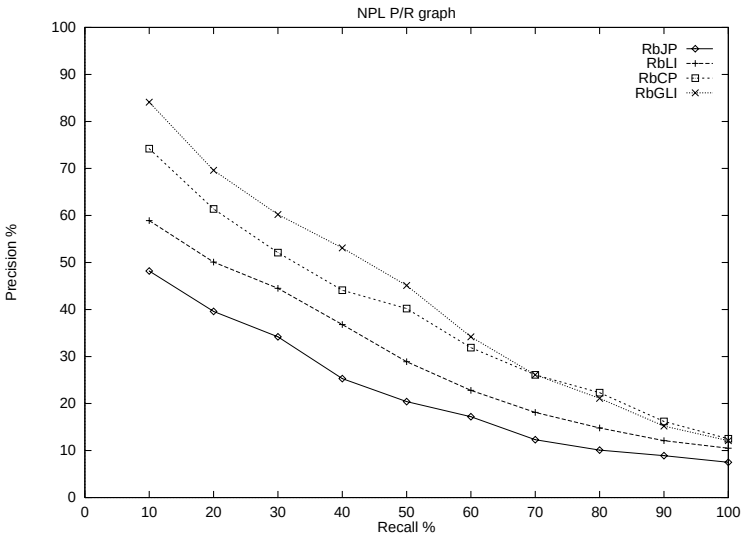

Figure 9: Precision and recall graph for the NPL test collection

8 Conclusions

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

Our current results, summarised in this paper, seem to suggest that an improvements in retrieval effectiveness can be obtained by designing IR systems that use probabilistic models based upon a different kind of probability kinematics. We intend to pursue this result to design a new probabilistic model for IR. We are currently exploring the possibilities of combining RbGLI with *Probabilistic Datalog*, a probabilistic extension of the Datalog logical model for Database systems proposed by Fuhr in [Fuhr, 1995]. Some initial results of this work can be found in [Rölleke, 1995, Crestani and Roelleke, 1995]. Moreover, we are trying to perform some further experiments on a much larger scale using a collection several

orders of magnitude larger than the ones used here. Some initial results of have been reported in [Crestani *et al.*, 1995].

Acknowledgements

The authors would like to thank the reviewers for their useful suggestions on how to improve the paper. Special thanks to Peter Bruza for the interesting discussions on the topics of this research, and to Mark Sanderson for his most valuable help in the experimentation. Finally, thanks to Juliet van Rijsbergen for her help in improving the readability of the paper.

References

[Amati and Kerpedjiev, 1992] G. Amati and S. Kerpedjiev. An Information Retrieval logical model: implementation and experiments. Technical Report Rel 5B04892, Fondazione Ugo Bordoni, Roma, Italy, March 1992.

[Brown *et al.*, 1992] P.F. Brown, V.J. Della Pietra, P.V. deSouza, J.C. Lai, and R.L. Mercer. Class-based n-gram models of natural language. *Computational Linguistics*, 18(4):467–479, 1992.

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

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

[Church and Hanks, 1989] K.W. Church and P. Hanks. Word association norms, mutual information and lexicography. In *Proccedings of ACL 27*, pages 76–83, Vancouver, Canada, 1989.

[Cleverdon *et al.*, 1966] C. Cleverdon, J. Mills, and M. Keen. *ASLIB Cranfield Research Project: factors determining the performance of indexing systems.* ASLIB, 1966.

[Cooper, 1971] W.S. Cooper. A definition of relevance for information retrieval. *Information Storage and Retrieval*, 7:19–37, 1971.

[Cooper, 1994] W.S. Cooper. The formalism of probability theory in IR: a foundation or an encumbrance. In *Proceedings of ACM SIGIR*, pages 242–247, Dublin, Ireland, June 1994.

[Crestani and Roelleke, 1995] F. Crestani and T. Roelleke. Issues on the implementation of imaging on top of probabilistic datalog. In *Proceedings of the First Workshop in IR, Uncertainty and Logic.* Glasgow, Scotland, UK, September 1995.

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

[Crestani *et al.*, 1995] F. Crestani, I. Ruthven, M. Sanderson, and C.J. van Rijsbergen. The troubles with using a logical model of IR on a large collection of documents. Experimenting retrieval by logical imaging on TREC. In *Proceedings of the Fourth Text Retrieval Conference (TREC-4)*, Washington D.C., USA, November 1995.

[Crestani *et al.*, 1996] F. Crestani, M. Sanderson, and C.J. van Rijsbergen. Sense resolution properties of logical imaging. *The New Review of Document and Text Management*, 1:277–298, 1996.

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

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

[Efthimiadis, 1996] E. Efthimiadis. Query expansion. *Annual Review of Information Science and Technology*, 31:121–187, 1996.

[Fuhr and Buckley, 1991] N. Fuhr and C. Buckley. A probabilistic learning approach for document indexing. *ACM Transactions on Information Systems*, 9(3):223–248, 1991.

[Fuhr, 1989] N. Fuhr. Models for retrieval with probabilistic indexing. *Information Processing and Management*, 25(1):55–72, 1989.

[Fuhr, 1992] N. Fuhr. Probabilistic models in Information Retrieval. *The Computer Journal*, 35(3):243–254, 1992.

[Fuhr, 1995] N. Fuhr. Probabilistic datalog - a logic for powerful retrieval methods. In *Proceedings of ACM SIGIR*, pages 282–290, Seattle, WA, USA, July 1995.

[Gärdenfors, 1982] P. Gärdenfors. Imaging and conditionalization. *Journal of Philosophy*, 79:747–760, 1982.

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

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

[Huibers *et al.*, 1996] T.W.C. Huibers, M. Lalmas, and C.J. van Rijsbergen. Information retrieval and situation theory. *SIGIR Forum*, 30(1):11–25, 1996.

[Kripke, 1971] 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.

[Lalmas and van Rijsbergen, 1993] M. Lalmas and C.J. van Rijsbergen. A model of an Information Retrieval system based on Situation Theory and Dempster-Shafer theory of evidence. In *Proceedings of the 1st Workshop on Incompleteness and Uncertainty in Information Systems*, pages 62–67, Montreal, Canada, 1993.

[Lewis, 1981] 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.

[Lewis, 1986] D. Lewis. *Conterfactuals*. Basil Blackwell, Oxford, UK, 2nd edition, 1986.

[Nie *et al.*, 1995] J.Y. Nie, F. Lepage, and M. Brisebois. Information retrieval as counterfactuals. *The Computer Journal*, 38(8):643–657, 1995.

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

[Nie, 1992] J.Y. Nie. Towards a probabilistic modal logic for semantic based information retrieval. In *Proceedings of ACM SIGIR*, pages 140–151, Copenhagen, Denmark, June 1992.

[Qiu and Frei, 1993] Y. Qiu and H.P. Frei. Concept based query expansion. In *Proceedings of ACM SIGIR*, pages 160–171, Pittsburgh, PA, USA, June 1993.

[Robertson and Sparck Jones, 1976] 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.

[Robertson, 1976] S.E. Robertson. *A theoretical model of the retrieval characteristics of information retrieval systems*. Phd thesis, University of London, UK, 1976.

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

[Rölleke, 1995] T. Rölleke. Imaging on top of probabilistic datalog. In *Proceedings of ACM SIGIR*, Seattle, WA, USA, July 1995. Poster session.

[Salton and Yang, 1973] G. Salton and C.S. Yang. On the specification of term values in automatic indexing. *Journal of Documentation*, 29(4):351–372, 1973.

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

[Sembok and van Rijsbergen, 1993] T.M.T Sembok and C.J. van Rijsbergen. Imaging: a relevance feedback retrieval with nearest neighbour clusters. In *Proceedings of the BCS Colloquium in Information Retrieval*, pages 91–107, Glasgow, UK, March 1993.

[Sparck Jones and van Rijsbergen, 1976] K. Sparck Jones and C.J. van Rijsbergen. Information Retrieval test collections. *Journal of Documentation*, 32(1):59–75, March 1976.

[Srinivadsan, 1992] P. Srinivadsan. Thesaurus construction. In W.B. Frakes and R. Baeza-Yates, editors, *Information Retrieval: data structures and algorithms.*, chapter 9. Prentice Hall, Englewood Cliffs, New Jersey, USA, 1992.

[Stalnaker, 1981] 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.

[van Rijsbergen, 1977] 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.

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

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

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

[Voorhees, 1993] E.M. Voorhees. On expanding query vectors with lexically related words. In *Proceeding of the TREC Conference*, pages 223–232, Gaithersburg, Maryland, USA, August 1993.

[Voorhees, 1994] E.M. Voorhees. Query expansion using lexical-semantic relations. In *Proceedings of ACM SIGIR*, pages 61–69, Dublin, Ireland, July 1994.

[Wong and Yao, 1989] S.K.M. Wong and Y.Y. Yao. A probability distribution model for Information Retrieval. *Information Processing and Management*, 25(1):39–53, 1989.

[Wong and Yao, 1991] S.K.M. Wong and Y.Y. Yao. A probabilistic inference model for information retrieval. *Information Systems*, 16(3):301–321, 1991.

[Wong and Yao, 1995] S.K.M. Wong and Y.Y. Yao. On modelling information retrieval with probabilistic inference. *ACM Transactions on Information Systems*, 13(1):38–68, 1995.

[Wong *et al.*, 1993] S.K.M. Wong, Y.J. Cai, and Y.Y. Yao. Computation of term association by a Neural Network. In *Proceedings of ACM SIGIR*, Pittsburgh, PA, USA, July 1993.