

“Is this document relevant? . . . probably” . A survey of probabilistic models in Information Retrieval

Fabio Crestani, Mounia Lalmas, Cornelis J. van Rijsbergen, and Iain Campbell
Computing Science Department
University of Glasgow

The paper provides an introduction to and survey of probabilistic approaches to modelling Information Retrieval. The basic concepts of probabilistic approaches to Information Retrieval are outlined, and the principles and assumptions upon which the approaches are based are presented. The various models that have been proposed in the development of IR are described, classified, and compared. The models are classified and compared using a common formalism. New approaches that constitute the basis of future research are described.

Categories and Subject Descriptors: H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval—*Retrieval Models*; H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing—*Indexing Methods*

General Terms: Algorithms, Theory

Additional Key Words and Phrases: Probabilistic Retrieval, Probabilistic Indexing, Probabilistic Modelling

1. HISTORY OF PROBABILISTIC MODELLING IN IR

In Information Retrieval (IR), probabilistic modelling is the use of a model that ranks documents in decreasing order of their evaluated probability of relevance to a user’s information need. Past and present research has made much use of formal theories of probability and of statistics in order to evaluate, or at least estimate, those probabilities of relevance. These attempts are to be distinguished from looser ones like, for example, the “vector space model” [Salton 1968] in which documents are ranked according to a measure of similarity with the query. A measure of similarity cannot be directly interpretable as a probability. In addition, similarity based models generally lack the theoretical soundness of probabilistic models.

The first attempts to develop a probabilistic theory of retrieval were made over thirty years ago [Maron and Kuhns 1960; Miller 1971]. Since, there has been a steady development of the approach. There are already several operational IR systems based upon probabilistic or semi-probabilistic models.

One major obstacle with probabilistic or semi-probabilistic IR models is that of finding methods for estimating the probabilities used to evaluate the probability of relevance that are both theoretically sound and computationally efficient. The problem of estimating these probabilities is difficult to tackle unless some simplify-

Address: Glasgow G12 8QQ, Scotland, UK, email: <fabio,mounia,keith,iain>@dcs.gla.ac.uk

ing assumptions are made. In the early stages of the study of probabilistic modelling in IR, assumptions related to event independence were employed in order to facilitate the computations. The first models to be based upon such assumptions were the “binary independence indexing model” (section 3.3) and the “binary independence retrieval model” (section 3.2). Recent findings by Cooper [Cooper 1995] have shown that these assumptions are not completely necessary and were, in fact, not actually made (section 5).

The earliest techniques that took into account dependencies gave results that were worse than those given by techniques based upon the simplifying assumptions. Moreover, the use of complex techniques that captured dependencies could only be made at a computational price regarded as too high with respect to the value of the results [van Rijsbergen 1977]. One particular research direction aimed at removing the simplifying assumptions has been studied extensively and much work is being done [Fung et al. 1990; Turtle and Croft 1990; Savoy 1992; van Rijsbergen 1992].

Another direction has involved the application of the statistical techniques used by pattern recognition and regression analysis. These investigations, of which the “Darmstadt indexing approach (DIA)” is a major example [Fuhr 1989; Fuhr and Buckley 1991] (see section 3.4), do not make use of independence assumptions. They are “model free” in the sense that the only probabilistic assumptions involved are those implicit in the statistical regression theory itself. The major drawback of such approaches is the degree to which heuristics are necessary to optimise the description and retrieval functions.

A theoretical improvement of the DIA was achieved through the use of logistic regression instead of standard regression. Standard regression is, strictly speaking, inappropriate for estimating probabilities of relevance where relevance is considered as a dichotomous event: i.e. a document is either relevant to a query or not. Logistic regression has been specifically developed to deal with dichotomous (or n-dichotomous) dependent variables. Probabilistic models that make use of logistic regression have been developed by Fuhr and Pfeifer in [Fuhr and Buckley 1991] and by Cooper et al. in [Cooper et al. 1992] (sections 3.4 and 3.7).

One area of recent research investigates the use of an explicit network representation of dependencies. The networks are processed by means of Bayesian inference or belief theory, using evidential reasoning techniques such as those described by Pearl [Pearl 1988]. This approach represents an extension of the earliest probabilistic models, taking into account the conditional dependencies present in a real environment. Moreover, the use of such networks generalises existing probabilistic models and allows the integration of several sources of evidence within a single framework. Attempts to use Bayesian (or causal) networks are reported in [Turtle 1990; Turtle and Croft 1991; Savoy 1992].

There is a new stream of research, initiated by van Rijsbergen [van Rijsbergen 1986] and continued by him and others [Amati and van Rijsbergen 1995; Lalmas 1996; Bruza 1993; Bruza and van der Weide 1992; Huibers 1996; Sebastiani 1994; Crestani and van Rijsbergen 1995]. It aims at developing a model based upon a non-classical logic, in particular, a conditional logic where the semantics are expressed using probability theory. The evaluation can be performed by means of a possible-world analysis [van Rijsbergen 1989; van Rijsbergen 1992; Sembok and van Rijsbergen 1993; Crestani and van Rijsbergen 1995] thus establishing an intentional

logic, by using modal logic [Nie 1988; Nie 1989; Amati and Kerpedjiev 1992; Nie 1992], by using situation theory [Lalmas 1992], or by integrating logic with Natural Language Processing [Chiaramella and Chevallet 1992]. The area is in its infancy; no working prototype based on the proposed models has been developed so far, and the operational validity of these ideas has still to be confirmed.

2. BACKGROUND

In this section, we review some general aspects that are important for a full understanding of the probabilistic models proposed so far. We then provide a framework within which the various models can be placed for comparison. We do not deal with concepts of probability theory in this paper. We assume some familiarity of principles of probability theory on the part of the reader. Finally, because of its importance to the foundations of all probabilistic retrieval models, we present the Probability Ranking Principle.

2.1 Event space

In general, probabilistic models have as their event space the set $\underline{Q} \times \underline{D}$, where \underline{Q} represents the set of all possible queries, and \underline{D} the set of all documents in the collection. The difference between the various models lies in their use of different *representations* and *descriptions* of queries and documents.

In most models, queries and documents are represented by descriptors, often automatically extracted or manually assigned terms. These descriptors are represented as binary valued vectors in which each element corresponds to a term. More complex models make use of real valued vectors, or take into account relationships among terms or among documents.

A *query* is an expression of an information need. In this paper, we regard a query as a unique event; that is, if two users submit the same query, or if the same query is submitted by the same user on two different occasions, these two queries are regarded as different queries. A query is submitted to the system, which then aims to find information *relevant* to the information need expressed in the query. In this paper we will consider relevance as a subjective user judgement on a document related to a unique expression of an information need¹.

A *document* is any object carrying information; a piece of text, an image, a sound, or a video. However, most all current IR systems deal only with text. This limitation results from problems associated with finding suitable representations for non textual objects. Therefore, in the remaining of this paper, we will consider only text-based IR systems.

Some assumptions common to all retrieval models:

—The users' understanding of their information need changes during a search session, is subject to a continuous refinement, and is expressed by different queries.

¹There exists a relevance relationship between a query and a document, which relies on a user perceived satisfaction of his or her information need. Such a perception of satisfaction is subjective - different users can give different relevance judgements to a given query-document pair. Moreover, this relevance relationship depends on time, so the same user could give a different relevance judgement on the same query-document pair on two different occasions.

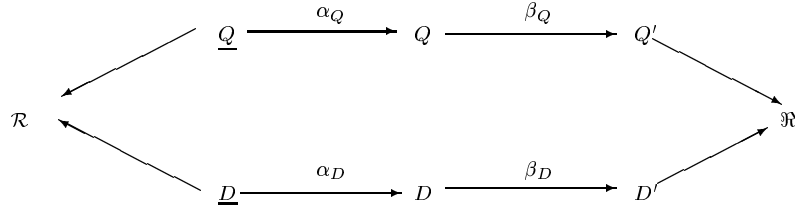


Fig. 1. The underlying conceptual model.

- Retrieval is based only upon representations of queries and documents, not upon the queries and documents themselves.
- The representation of IR objects is “uncertain”. For example, the extraction of index terms from a document or a query to represent the document or query informative content is a highly uncertain process. As a consequence, the retrieval process becomes uncertain.

It is particularly this last assumption that gave way to the study of probabilistic retrieval models. Probability theory [Good 1950] is, however, only one way of dealing with uncertainty². Also, earlier models were largely based on classical probability theory, but, in recent times new approaches to dealing with uncertainty have been applied to IR. Sections 3 and 4 present both traditional and new approaches to probabilistic retrieval.

2.2 A conceptual model

The importance of conceptual modelling is widely recognised in fields such as Database Management Systems and Information Systems. For this survey, we will use the conceptual model proposed by Fuhr [Fuhr 1992b], which has the advantage of being both simple and general enough to be considered a conceptual basis for all the probabilistic models presented in this survey, although some of them predate it.

The model is shown in Figure 1. The basic objects of an IR system are: a finite set of *documents* \underline{D} (e.g., books, articles, images) and a finite set of *queries* \underline{Q} (e.g., information needs). We consider a set of queries and not a single query alone because a single user may have varying information needs. If we consider \mathcal{R} a finite set of possible relevance judgements, for example in the binary case $\mathcal{R} = \{R, \bar{R}\}$, that is, a document can either be relevant or not to a query, then the IR system’s task is to map every query-document pair to an element of \mathcal{R} . Unfortunately, IR systems do not deal directly with queries and documents, but with *representations* of them (e.g., a text for a document, or a Boolean expression for a query). It is

²Other approaches are based, for example, on Fuzzy Logic [Zadeh 1987] and Dempster-Shafer’s theory of evidence [Shafer 1976].

mainly the kind of representation technique used that differentiates one IR model from another.

We denote α_Q the mapping between a set of queries \underline{Q} and their representations Q . For example, a user in search of information about wine may express his or her query as follows: “I am looking for articles dealing with wine”. Similarly, we denote α_D the mapping between a set of documents \underline{D} and their representations D . For example, in a library, a book is represented by its author, titles, a summary, the fact it is a book (and not a article), and some keywords. These two mappings can be very different from each other. Obviously, the better the representation of queries and documents, the better will be the performance of the IR system.

To make the conceptual model general enough to deal with the most complex IR models, a further mapping has been introduced between representations and *descriptions*. For instance, a description of the above query could be the following two stems: “article” and “wine”. The sets of representations Q and D are mapped to the sets of descriptions Q' and D' by means of two mapping functions β_Q and β_D . Moreover, the need for such additional mapping arises for learning models (see for example section 3.4) that have to aggregate features to allow large enough samples for estimation. It is worth noticing, however, that most models work directly with the original document and query representations.

It is common for IR systems to be able to manage only a poor description of the representation of the objects (e.g., a set of stems instead of a text). However, when representation and description happen to be the same, it is sufficient to consider either α_Q or α_D as an identity mapping.

Descriptions are taken as the independent variables of the retrieval function $r : Q' \times D' \rightarrow \mathfrak{R}$, which maps query-document pair onto a set of *retrieval status values* (RSV) $r(q'_k, d'_j)$ [Bookstein and Cooper 1976]. The task of ranked retrieval IR systems in response to a query \underline{q}_k is to calculate this value and rank each and every document \underline{d}_j in the collection upon it.

In probabilistic IR the task of the system is different. If we assume binary relevance judgements, i.e. \mathcal{R} contains only the two possible judgements R and \bar{R} , then according to the Probability Ranking Principle (section 2.4), the task of an IR system is to rank the documents according to their estimated probability of being relevant $P(R | \underline{q}_k, \underline{d}_j)$. This probability is estimated by $P(R | q'_k, d'_j)$, which is the retrieval status value.

2.3 On the concepts of “relevance” and “probability of relevance”

The concept of *relevance* is arguably the fundamental concept of IR. In the above presented model we purposely avoid giving a formal definition of relevance. The reason behind our decision is that the notion of relevance has never been defined precisely in IR. Although there has been a large number of attempts towards a definition of the concept of relevance [Seracevic 1970; Cooper 1971; Mizzaro 1996], there has never been agreement about unique precise definition. A treatment of the concept of relevance is outside the scope of this paper and we will not attempt to formulate a new definition or even accept a particular already existing one. What is important for the purpose of our survey is to understand that relevance is a relationship that may or may not hold between a document and a user of the IR system who is searching for some information: if the user wants the document in

question, then we say that the relationship holds. With reference to the model presented above, relevance (\mathcal{R}) is a relationship between a document (\underline{d}_j) and a user's information need (q_k). If the user wants the document \underline{d} in relation to his information need q_k , then \underline{d}_j is relevant (R).

Most readers will find the concept of *probability of relevance* quite unusual. The necessity of introducing such probability arises from the fact that relevance is a function of a large number of variables concerning the document, the user, and the information need. It is virtually impossible to make strict prediction as to whether the relationship of relevance will hold between a given document and a given user's information need. The problem must be approached probabilistically. The above model explains what is the evidence available to an IR system to estimate the probability of relevance $P(R | q_k, \underline{d}_j)$. A precise definition of probability of relevance depends on a precise definition of the concept of relevance, and given a precise definition of relevance it is possible to define rigorously such probability. Just as we did not define relevance, we will not attempt to define the probability of relevance, since every model presented here uses a somewhat different definition. We refer the reader to the treatment given by Robertson et al. in [Robertson et al. 1982], where different interpretations of the probability of relevance are given and a unified view is proposed.

2.4 The Probability Ranking Principle

A common characteristic of all the probabilistic models developed in IR is their adherence to the theoretical justification embodied in the Probability Ranking Principle (PRP) [Robertson 1977]. The PRP asserts that optimal retrieval performance can be achieved when documents are ranked according to their probabilities of being judged relevant to a query. The above probabilities should be estimated as accurately as possible on the basis of whatever data has been made available for this purpose.

The principle speaks of “optimal retrieval”, as distinct from “perfect retrieval”. Optimal retrieval can be defined precisely for probabilistic IR because it can be proved theoretically with respect to representations (or descriptions) of documents and information needs. Perfect retrieval relates to the objects of the IR systems themselves, i.e., documents and information needs.

The formal definition of the PRP is as follows. Let C denote the cost retrieving a relevant document, and \overline{C} the cost of retrieving a non-relevant document. The decision rule that is the basis of the PRP states that a document d_m should be retrieved in response to a query q_k above any document d_i in the collection if:

$$C \cdot P(R | q_k, d_m) + \overline{C} \cdot (1 - P(R | q_k, d_m)) \leq P(R | q_k, d_i) + \overline{C} \cdot (1 - P(R | q_k, d_i))$$

The decision rule can be extended to deal with multi-valued relevance scales, (e.g., very relevant, possibly relevant, etc. [Cooper 1971]). In addition, by means of a continuous cost function, it is possible to write a decision rule for approaches where the relevance scale is assumed to be continuous [Borgogna and Pasi 1993].

The application of the PRP in probabilistic models involves assumptions:

—Dependencies between documents are generally ignored. Documents are considered in isolation, so that the relevance of one document to a query is considered

independent from that of other documents in the collection (nevertheless, see section 5).

- It is assumed that the probabilities (e.g., $P(R | q_k, d_i)$) in the decision function can be estimated in the best possible way, that is accurately enough to approximate the user's real relevance judgement, and therefore order the documents accordingly.

Although these assumptions limit the applicability of the PRP, models based on it enable the implementation of IR systems offering some of highest level of retrieval performance currently available [Robertson 1977]. There are, of course, a number of other retrieval strategies with high levels of performance and that are not consistent with the PRP. Examples of such strategies are the Boolean or the cluster model. In this paper we are not concerned with these models since they are not probabilistic in nature and do not fall into the class of models this survey is about.

2.5 The remainder of this paper

In the remainder of this paper, we present a survey of probabilistic IR models in two main categories: *relevance models* and *inference models*.

Relevance models are described in section 3. These models are based on evidence about which documents are relevant to a given query. The problem of estimating the probability of relevance for every document in the collection is difficult because of the large number of variables involved in the representation of documents in comparison to the small amount of document relevance information available. The models differ, primarily, in the way they estimate this or related probabilities.

Inference models are presented in section 4. These models apply concepts and techniques originating from areas such as logic and artificial intelligence. From a probabilistic perspective, the most noteworthy examples are those that consider IR as process of uncertain inference. The concept of relevance is interpreted in a different way, where it can be extended and defined with respect, not only to a query formulation, but also to an information need.

The models of both categories are presented separately, but using a common formalism and, as much as possible, to the same level of detail.

We would also like to point out that in this paper we are not concerned with issues related to evaluation. Evaluation is a very important part of IR research and even a brief treatment of some the issues involved in the area would require an entire paper. We suggest the interested reader to look at the extensive IR literature on this subject, and in particular at [van Rijsbergen 1979; Sparck Jones 1981].

3. PROBABILISTIC RELEVANCE MODELS

The main task of IR systems based upon relevance models is to evaluate a probability of a document being relevant. This is done by estimating the probability $P(R | q_k, d_i)$ for every document d_i in the collection, which is a difficult problem. The estimation problem can only be tackled by means of simplifying assumptions. Two kinds of approaches have been developed to deal with such assumptions: model-oriented and description-oriented.

Model-oriented approaches are based upon some probabilistic independence assumptions concerning the elements used in representing³ the documents or the queries. The probabilities of these individual representation elements are estimated, and, by means of the independence assumptions, the probabilities of the document representations are estimated from them. The Binary Independence Indexing and Retrieval models (sections 3.3 and 3.2), and the n-Poisson model (section 3.8) are examples of this approach.

Description-oriented approaches are more heuristic in nature. Given the representation of queries and documents, a set of features for query-document pairs is defined (e.g., occurrence frequency information), that allows each query-document pair in the collection to be mapped on to these features. Then, by means of some training data containing query-document pairs together with their corresponding relevance judgements, the probability of relevance is estimated with respect to these features. The best example of the application of this approach is the Darmstadt Indexing model (section 3.4). However, a new model whose experimental results are not yet known, has been proposed by Cooper et al. [Cooper et al. 1992]. These models exploit the mapping between representations and descriptions that we introduced in section 2.2.

3.1 Probabilistic Modelling as a decision strategy

The use of probabilities in IR was advanced in 1960 by Maron and Kuhns [Maron and Kuhns 1960]. In 1976, Robertson and Sparck Jones went further by showing the powerful contribution of probability theory to model IR. The probabilistic model was theoretically finalised by van Rijsbergen in [van Rijsbergen 1979], chapter 6. The focus of the model is on its analysis as a decision strategy based upon a loss or risk function.

Referring to the conceptual model described in section 2.2, it is assumed that the representation and the description methods for queries and documents are the same. Queries and documents are described by sets of index terms. Let $T = \{t_1, \dots, t_n\}$ denote the set of terms used in the collection of documents. We represent the query q_k with terms belonging to T . Similarly, we represent a document d_j as the set of terms occurring in it. If we use a binary representation then d_j is represented as the binary vector $\vec{x} = (x_1, \dots, x_n)$ with $x_i = 1$ if $t_i \in d_j$ and $x_i = 0$ otherwise. The query q_k is represented in the same manner.

The basic assumption, common to most models described in section 3, is that the distribution of terms within the document collection provides information concerning the relevance of a document to a given query. This is because it is assumed that terms are distributed differently in relevant and non-relevant documents. This is known as the *cluster hypothesis* (see [van Rijsbergen 1979] pp. 45-47). If the term distribution was the same within the sets of relevant and non-relevant documents then it would not be possible to devise a discrimination criterion between them. In which case, a different representation of the document information content would be necessary.

³Depending on the complexity of the models, the probabilities to be estimated can be with respect to the representations or the descriptions. But for clarity of expression, we will refer to the representations only, unless otherwise stated.

$C_j(R, dec)$	retrieved	not retrieved
relevant document	0	λ_1
non relevant document	λ_2	0

Table 1. The cost of retrieving and not retrieving a relevant and non relevant document

The term distribution provides information about the “probability of relevance” of a document to a query. If we assume binary relevance judgements, then the term distribution provides information about $P(R | q_k, d_j)$.

The quantity $P(R | q_k, \vec{x})$, with \vec{x} as a binary document representation, cannot be estimated directly. Instead, Bayes’ theorem is applied [Pearl 1988]:

$$P(R | q_k, \vec{x}) = \frac{P(R | q_k) \cdot P(\vec{x} | R, q_k)}{P(\vec{x} | q_k)}$$

To simplify notation, we omit the q_k on the understanding that evaluations are with respect to a given query q_k . The previous relation becomes:

$$P(R | \vec{x}) = \frac{P(R) \cdot P(\vec{x} | R)}{P(\vec{x})}$$

where $P(R)$ is the prior probability of relevance, $P(\vec{x} | R)$ is the probability of observing the description \vec{x} conditioned upon relevance having been observed, and $P(\vec{x})$ is the probability that x is observed. The latter is determined as the joint probability distribution of the n terms within the collection. The above formula evaluates the “posterior” probability of relevance conditioned upon the information provided in the vector \vec{x} .

The provision of a ranking of documents by the PRP can be extended to provide an “optimal threshold” value. This can be used to set a cut-off point in the ranking to distinguish between those documents that are worth retrieving and those that are not. This threshold is determined by means of a *decision strategy*, whose associated *cost function* $C_j(R, dec)$ for each document d_j is described in Table 1.

The decision strategy can be described simply as one that minimises the average cost resulting from any decision. This strategy is equivalent to minimising the following *risk function*:

$$\mathcal{R}(R, dec) = \sum_{d_j \in D} C_j(R, dec) \cdot P(d_j | R)$$

It can be shown (see [van Rijsbergen 1979], pp. 115-117) that the minimisation of that function brings about an optimal partitioning of the document collection. This is achieved by retrieving only those documents for which the following relation holds:

$$\frac{P(d_j | R)}{P(d_j | \bar{R})} > \lambda$$

where

$$\lambda = \frac{\lambda_2 \cdot P(\bar{R})}{\lambda_1 \cdot P(R)}$$

3.2 The Binary Independence Retrieval model

In the previous section, it remains necessary to estimate the joint probabilities $P(d_j | R)$ and $P(d_j | \overline{R})$, that is $P(\vec{x} | R)$ and $P(\vec{x} | \overline{R})$ if we consider the binary vector document representation \vec{x} .

In order to simplify the estimation process, the components of the vector \vec{x} are assumed to be stochastically independent when conditionally dependent upon R or \overline{R} . That is, the joint probability distribution of the terms in the document d_j is given by the following product of marginal probability distributions:

$$P(d_j | R) = P(\vec{x} | R) = \prod_{i=1}^n P(x_i | R)$$

and

$$P(d_j | \overline{R}) = P(\vec{x} | \overline{R}) = \prod_{i=1}^n P(x_i | \overline{R})$$

This *binary independence assumption*, is the basis of a model first proposed by Robertson and Spark Jones in 1976 [Robertson and Sparck Jones 1976]: the *Binary Independence Retrieval model* (BIR). The assumption has always been recognised as unrealistic.

Nevertheless, as pointed out by Cooper (section 5), the assumption that actually underpins the BIR model is not that of binary independence, but that of the weaker assumption of *linked dependence*:

$$\frac{P(\vec{x} | R)}{P(\vec{x} | \overline{R})} = \prod_{i=1}^n \frac{P(x_i | R)}{P(x_i | \overline{R})}$$

This states that the ratio between the probabilities of \vec{x} occurring in relevant and non relevant documents is equal to the product of the corresponding ratios of the single terms.

Considering the decision strategy of the previous section, it is now possible to obtain a decision strategy by using a logarithmic transformation to obtain a linear decision function:

$$g(d_j) = \log \frac{P(d_j | R)}{P(d_j | \overline{R})} > \log \lambda$$

To simplify notation, we define the following quantities: $p_j = P(x_j = 1 | R)$, and $q_j = P(x_j = 1 | \overline{R})$ which represent the probability of the j th term appearing in a relevant, and in a non relevant document, respectively. Clearly: $1 - p_j = P(x_j = 0 | R)$, and $1 - q_j = P(x_j = 0 | \overline{R})$. This gives:

$$P(\vec{x} | R) = \prod_{j=1}^n p_j^{x_j} \cdot (1 - p_j)^{1-x_j}$$

and

$$P(\vec{x} | \overline{R}) = \prod_{j=1}^n q_j^{x_j} \cdot (1 - q_j)^{1-x_j}$$

Substituting the above, gives:

$$g(d_i) = \sum_{j=1}^n (x_j \cdot \log \frac{p_j}{q_j} + (1 - x_j) \cdot \log \frac{1-p_j}{1-q_j}) \\ = \sum_{j=1}^n c_j x_j + C$$

where:

$$c_j = \log \frac{p_j \cdot (1 - q_j)}{q_j \cdot (1 - p_j)}$$

and

$$C = \sum_{j=1}^n \log \frac{1 - p_j}{1 - q_j}$$

This formula gives the RSV of document d_j for the query under consideration. Documents are ranked according to their RSV and presented to the user. The cut-off value λ can be used to determine the point at which the display of the documents is stopped, although, the RSV is generally used only to rank the entire collection of documents. In a real IR system, the presentation of documents ordered on their estimated probability of relevance to a query matters more than the actual value of those probabilities. Therefore, since the value of C is constant for a specific query, we need only consider the value of c_j . This value, or more often the value $\exp(c_j)$, is called the *term relevance weight* (TRW), and indicates the term's capability to discriminate relevant from non relevant documents. As it can be seen, in the BIR model term relevance weights contribute "independently" to the relevance of a document.

To apply the BIR model, it is necessary to estimate the parameters p_j and q_j for each term used in the query. This is performed in various ways, depending upon the amount of information available. The estimation can be retrospective or predictive. The first is used on test collections where the relevance assessments are known. The second is used with normal collection where parameters are estimated by means of relevance feedback from the user.

There is another technique, proposed by Croft and Harper [Croft and Harper 1979], that uses a collection information to make estimates and does not use relevance information. Let us assume that the IR system has already retrieved some documents for the query q_k . The user is asked to give relevance assessments for those documents, from which the parameters of the BIR are estimated. If we also assume to be working in the retrospective case, then we know the relevance value of all individual documents in the collection. Let a collection have N documents, R of which are relevant to the query. Let n_j denote the number of documents in which the term x_j appears, amongst which, only r_j are relevant to the query. The parameters p_j and q_j can then be estimated as follows:

$$\hat{p}_j = \frac{r_j}{R}$$

and

$$\hat{q}_j = \frac{n_j - r_j}{N - R}$$

These give:

$$TRW_j = \frac{\frac{r_j}{R-r_j}}{\frac{n_j-r_j}{N-n_j-R+r_j}}$$

This approach is possible only if we have relevance assessments for all documents in the collection, i.e. where we know R and r_j . According to Croft and Harper, given that the only information concerning the relevance of documents is that provided by a user through relevance feedback, predictive estimations should be used. Let \tilde{R} denote the number of documents judged relevant by the user. Further, let \tilde{r}_j be the number of those documents in which the term x_j occurs. We can then combine this with the estimation technique of [Cox 1970].

$$T\tilde{R}W_j = \frac{\frac{\tilde{r}_j+0.5}{R-\tilde{r}_j+0.5}}{\frac{n_j-\tilde{r}_j+0.5}{N-n_j-R+\tilde{r}_j+0.5}}$$

Usually, the relevance information given by a user is limited and is not sufficiently representative of the entire collection. Consequently, the resulting estimates tend to lack precision. As a partial solution, one generally simplifies by assuming p_j to be constant for all the terms in the indexing vocabulary. The value $p_j = 0.5$ is often used, which gives a TRW that can be evaluated easily:

$$T\tilde{R}W_j = \frac{N-n_j}{n_j}$$

For large N , i.e. large collections of documents, this expression can be approximated by the “inverse document frequency” $IDF_j = \log N/n_j$. This is widely used in IR to provide an intuitive discrimination power of a term in a document collection.

3.3 The Binary Independence Indexing model

The *Binary Independence Indexing model* (BII model) is a variant of the BIR model. Where the BIR model regards a single query with respect to the entire document collection, the BII model regards one document in relation to a number of queries. The indexing weight of a term is evaluated as an estimate of the probability of relevance of that document with respect to queries using that term. This idea was first proposed in Maron and Kuhns’s indexing model [Maron and Kuhns 1960].

In the BII, the focus is on the query representation, which we assume to be a binary vector \vec{z} . The dimension of the vector is given by the set of all terms T which could be used in a query, and $z_j = 1$ if the term represented by that element is present in the query, $z_j = 0$ otherwise⁴. In this model, the terms weights are defined in terms of frequency information derived from queries; that is, an explicit document representation is not required. We will only assume that there is a subset of terms that can be used to represent any document, and that will be given weights with respect to a particular document.

⁴As a consequence, two different information needs (i.e., two queries) using the same set of terms will produce the same ranking of documents.

The BII model seeks an estimate of the probability $P(R | \vec{z}, d_j)$ that the document d_j will be judged relevant to the query represented by \vec{z} . To use the same formalism as the previous section, we use \vec{x} to denote the document representation. So far this model looks very similar to the BIR; the difference lies with the application of Bayes' theorem as follows:

$$P(R | \vec{z}, \vec{x}) = \frac{P(R | \vec{x}) \cdot P(\vec{z} | R, \vec{x})}{P(\vec{z} | \vec{x})}$$

$P(R | \vec{x})$ is the probability that the documents represented by \vec{x} will be judged relevant to an arbitrary query. $P(\vec{z} | R, \vec{x})$ is the probability that the document will be relevant to a query with representation \vec{z} . As \vec{z} and \vec{x} are assumed to be mutually independent, $P(\vec{z} | \vec{x})$ reduces to the probability that the query \vec{z} will be submitted to the system $P(\vec{z})$.

To proceed from here, some simplifying assumptions must be made:

- (1) The conditional distribution of terms in all queries is independent. This is the classic "binary independence assumption", from which the model's name arises:

$$P(\vec{z} | R, \vec{x}) = \prod_{i=1}^n P(z_i | R, \vec{x})$$

- (2) The relevance of a document with representation \vec{x} with respect to a query \vec{z} depends only upon the terms used by the query (i.e., those with $z_i = 1$) and not upon other terms.
- (3) With respect to a specific document, for each term not used in the document representation, we assume:

$$P(R | z_i, \vec{x}) = P(R | \vec{x})$$

Now, applying the first assumption to $P(R | \vec{z}, \vec{x})$, we get:

$$P(R | \vec{z}, \vec{x}) = \frac{P(R | \vec{x})}{P(\vec{z} | \vec{x})} \cdot \prod_{i=1}^n P(z_i | R, \vec{x})$$

by applying the second assumption and Bayes' theorem, we get the ranking formula:

$$\begin{aligned} P(R | \vec{z}, \vec{x}) &= \frac{\prod_i P(z_i)}{P(\vec{z})} \cdot \prod_{i=1}^n \frac{P(R|z_i, \vec{x})}{P(R|\vec{x})} \\ &= \frac{\prod_i P(z_i)}{P(\vec{z})} \cdot P(R | \vec{x}) \cdot \prod_{z_i=1} \frac{P(R|z_i=1, \vec{x})}{P(R|\vec{x})} \cdot \prod_{z_i=0} \frac{P(R|z_i=0, \vec{x})}{P(R|\vec{x})} \end{aligned}$$

The value of the first fraction is a constant c for a given query, so there is no need to estimate it for ranking purposes. In addition, by applying the third assumption, the third fraction becomes equal to 1, and we obtain:

$$P(R | \vec{z}, \vec{x}) = c \cdot P(R | \vec{x}) \cdot \prod_{t_i \in \vec{z} \cap \vec{x}} \frac{P(R | t_i, \vec{x})}{P(R | \vec{x})}$$

There are a few problems with this model. The use of the third assumption is in contrast with experimental results reported by Turtle [Turtle 1990], who demonstrates the advantage of assigning weights to query terms not occurring in a document. Moreover, the second assumption is called into question by Robertson et al.

[Robertson and Sparck Jones 1976]. They proved experimentally the superiority of a ranking approach in which the probability of relevance is based upon both the presence and the absence of query terms in documents. The results suggest that the BII model might obtain better results if it were, for example, used together with a thesaurus or a set of term-term relations. This would enable the use of document terms not present in the query, but related in some way to those that were.

Fuhr [Fuhr 1992b] pointed out that, in its present form, the BII model is hardly an appropriate model because, in general, there is not enough relevance information available to estimate the probability $P(R | t_i, \vec{x})$ for specific term-document pairs. To partially overcome this problem, one can assume that a document consists of independent components to which the indexing weights relate. However, experimental evaluations of this strategy have shown only average retrieval results [Kwok 1990].

Robertson et al. proposed a model that provides a unification of the BII and BIR models [Robertson et al. 1982]. The proposed model, simply called Model 3 (as opposed to the BII model called Model 1 and the BIR model called Model 2), enables to combine the two retrieval strategies of the BII and the BIR models, thus providing a new definition of probability of relevance that unifies those of the BII and BIR models. In the BII model the probability of relevance of a document given a query is computed relative to evidence consisting of the properties of the queries for which that document was considered relevant, while in the BIR model it is computed relative to the evidence consisting of the properties of documents considered relevant by that same query. Model 3 enables to use both forms of evidence. Unfortunately, a computationally treatable estimation theory fully faithful to Model 3 has not been proposed. The Model 3 idea has been explored later by Fuhr [Fuhr 1989] and Wong and Yao [Wong and Yao 1989] (see section 3.5).

3.4 The Darmstadt Indexing model

The basic idea of the *Darmstadt Indexing approach* (DIA) is to use long-term learning of indexing weights from users' relevance judgements [Fuhr and Knowrz 1984; Biebricher et al. 1988; Fuhr and Buckley 1991]. It can be seen as an attempt to develop index term specific estimates based upon the use of index terms in the learning sample.

DIA attempts to estimate $P(R | x_i, q_k)$ from a sample of relevance judgements of query-document or term-document pairs. This approach, when used for indexing, associates a set of heuristically selected attributes to each term-document pair, rather than estimating the probability associated with an index term directly (examples are given below). The use of an attribute set reduces the amount of training data required and allows the learning to be collection specific. However, the degree to which the resulting estimates are term specific depends critically upon the particular attributes used.

The indexing performed by the DIA is divided in two steps: a description step and a decision step.

In the *description step* relevance descriptions for term-document pairs (x_i, \vec{x}) are formed. These relevance descriptions $s(x_i, \vec{x})$ ⁵, comprise a set of attributes

⁵These are similar to those used in pattern recognition.

considered important for the task of assigning weights to terms with respect to documents. A relevance description $s(x_i, \vec{x})$ contains values of attributes of the term x_i , of the document (represented by \vec{x}) and of their relationships. This approach does not make any assumptions about the structure of the function s or about the choice of attributes. Some examples of attributes which could be used by the description function are:

- frequency of occurrence of term x_i in the document,
- inverse frequency of term x_i in the collection,
- information about the location of the occurrence of term x_i in the document, or
- parameters describing the document, e.g. its length, the number of different terms occurring in it, etc.

In the *decision step*, a probabilistic index weight based on the previous data is assigned. This means that we estimate $P(R | s(x_i, \vec{x}))$ and not $P(R | x_i, \vec{x})$. In the latter case, we would have regarded a single document d_j (or \vec{x}) with respect to all queries containing x_i , as in the BII model. Here, we regard the set of all query-document pairs in which the same relevance description s occurs. The interpretation of $P(R | s(x_i, \vec{x}))$ is therefore that of the probability of a document being judged relevant to an arbitrary query, given that a term common to both document and query has a relevance description $s(x_i, \vec{x})$.

The estimates of $P(R | s(x_i, \vec{x}))$ are derived from a learning sample of term-document pairs with attached relevance judgements derived from the query-document pairs. If we call this new domain L , we have:

$$L \subset \underline{D}xQx\mathfrak{R}, \text{ or } L = \{(q_k, d_j, r_{kj})\}$$

By forming relevance descriptions for the terms common to queries and documents for every query-document pair in L , we get a multi-set of relevance descriptions with relevance judgements:

$$L^x = [(s(x_i, d_j), r_{kj}) | x_i \in q_k \cap d_j \wedge (q_k, d_j, r_{kj}) \in L]$$

Using this set, it would be possible to estimate $P(R | s(x_i, \vec{x}))$ as the relative frequency of those elements of L^x with the same relevance description. Nevertheless, the technique used in DIA makes use of an *indexing function*, because it provides better estimates through the use of additional plausible assumptions about the indexing function. In [Fuhr and Buckley 1991], various linear indexing functions estimated by least squares polynomial were used, while in [Fuhr and Buckley 1993] a logistic indexing function estimated by maximum likelihood was attempted. Experiments were performed using both a controlled and a free term vocabulary.

The experimental results on the standard test collections indicate that the DIA approach is often superior to other indexing methods. The more recent, but only partial, results obtained using the TREC collection [Fuhr and Buckley 1993] tend to support this conclusion.

3.5 The Retrieval with Probabilistic Indexing model

The *Retrieval with Probabilistic Indexing* (RPI) model described in [Fuhr 1989] takes a different approach from other probabilistic models. This model assumes

that we use not only a weighting of index terms with respect to the document but also a weighting of query terms with respect to the query. If we denote w_{mi} the weight of index term x_i with respect to the document \vec{x}_m , and v_{ki} the weight of query term $z_i = x_i$ with regard to the query \vec{z}_k , then we can evaluate the following scalar product and use it as retrieval function:

$$r(\vec{x}_m, \vec{z}_k) = \sum_{\{x_m=z_k\}} w_{mi} \cdot v_{ki}$$

Wong and Yao [Wong and Yao 1989] give an utility theoretic interpretation of this formula for probabilistic indexing. Assuming we have a weighting of terms with respect to documents (similar to those, for example, of BII or DIA), the weight v_{ki} can be regarded as the utility of the term t_i , and the retrieval function $r(d_m, q_k)$ as the expected utility of the document with respect to the query. Therefore, $r(d_m, q_k)$ does not estimate the probability of relevance, but it has the same utility theoretic justification as the PRP.

RPI was developed especially for combining probabilistic indexing weighting with query term weighting based, for example, on relevance feedback. As a result, its main advantage is that it is suitable for application to different probabilistic indexing schemes.

3.6 The Probabilistic Inference model

Wong and Yao in [Wong and Yao 1995] extend the work reported in [Wong and Yao 1989] by using an epistemological view of probability, from where they proposed a probabilistic inference model for IR. With the epistemic view of probability theory, the probabilities under consideration are defined based on semantic relationships between documents and queries. The probabilities are interpreted as degrees of beliefs.

The general idea of the model starts with the definition of a concept space, which can be interpreted as the knowledge space in which documents, index terms, and user queries are represented as propositions. For example: the proposition d is the knowledge contained in the document; the proposition q is the information need requested; and the proposition $d \cap q$ is the portion of knowledge common to d and q .

An epistemic probability function P is defined on the concept space. For example, $P(d)$ is the degree to which the concept space is covered by the knowledge contained in the document and $P(d \cap q)$ is the degree to which the concept space is covered by the knowledge common to the document and the query.

Based on these probabilities, different measures can be constructed to evaluate the relevance of documents to queries, offering different interpretations of relevance, thus leading to different approaches to model IR. We discuss two of them. The first one is:

$$\Psi(d \rightarrow q) = P(q|d) = \frac{P(d \cap q)}{P(d)}$$

$\Psi(d \rightarrow q)$ can be considered as a measure of precision of the document with respect to the query, and is defined as the probability that a retrieved document is relevant. A precision-oriented interpretation of relevance should be used when a

user is interested in locating a specific piece of information. A second measure is:

$$\Psi(q \rightarrow d) = P(d|q) = \frac{P(q \cap d)}{P(q)}$$

$\Psi(q \rightarrow d)$ is considered as a recall index of the document with respect to the query, and is defined as the probability that a relevant document is retrieved. A recall-oriented measure should be used when the user is writing a review paper on a particular subject, and is interested in finding as much papers as possible on the subject.

Depending of the relationships between concepts, different formulations of $\Psi(d \rightarrow q)$ and $\Psi(q \rightarrow d)$ are obtained. For example, suppose that the concept space is $t_1 \cup \dots \cup t_n$ where the basic concepts are (pairwise) disjoint; i.e., $t_i \cap t_j = \emptyset$ for $i \neq j$. It can be proven that:

$$\Psi(d \rightarrow q) = \frac{\sum_t P(d \cap q|t)P(t)}{P(d)}$$

$$\Psi(q \rightarrow d) = \frac{\sum_t P(d \cap q|t)P(t)}{P(q)}$$

Wong and Yao work aims to provide a probabilistic evaluation of uncertain implications which have been advanced as a way to measure the relevance of documents to queries (see section 4.1). Although measuring uncertain implications by a probability function is more restrictive than for example using the possible world analysis, the model proposed by Wong and Yao is both expressive and sound. For example, they show that the Boolean, fuzzy set, vector space and probabilistic models are special cases of their model. We will not go into the detail of this demonstration, but we refer to the cited articles.

3.7 The Staged Logistic Regression model

Cooper's *Staged Logistic Regression model* (SLR), proposed in [Cooper et al. 1992], is an attempt to overcome some problems present in the use of standard regression methods to estimate probabilities of relevance in IR. Cooper criticises Fuhr's approaches, especially the DIA which require strong simplifying assumptions. He thinks (we include a longer explanation of his point of view in section 5) that these assumptions inevitably distort the final estimate of the probability of relevance. He advocates a "model-free" approach to estimation. In addition, a more serious problem lies in the use of standard polynomial regression methods. Standard regression theory is based on the assumption that the sample values taken for the dependent variable are from a continuum of possible magnitudes. In IR, the dependent variable is usually dichotomous: a document is either relevant or non relevant. So standard regression is clearly inappropriate in such cases.

A more appropriate tool, according to Cooper, is *logistic regression*, a statistical method specifically developed for using dichotomous (or discrete) dependent variables. Related techniques were used with some success by other researchers, for example, Fuhr employed it in [Fuhr and Pfeifer 1991] and more recently in [Fuhr and Buckley 1993].

The method proposed by Cooper is based on the guiding notion of treating composite clues on at least two levels, an intra-clue level at which a predictive

statistic is estimated separately for each composite clue⁶, and an inter-clue level in which these separate statistics are combined to obtain an estimate of the probability of relevance for a query-document pair. As this proceeds in stages, the method is called Staged Logistic Regression (SLR). A two stage SLR would be as follows:

- (1) A statistical simplifying assumption is used to break down the complex joint probabilistic distribution of the composite clues. This assumption is called *linked dependence*. For example, assuming that we have only two clues, a positive real number K exists such that the following conditions hold true:

$$P(a, b | R) = K P(a | R) \cdot P(b | R)$$

$$P(a, b | \neg R) = K P(a | \neg R) \cdot P(b | \neg R)$$

It follows that:

$$\frac{P(a, b | R)}{P(a, b | \neg R)} = \frac{P(a | R)}{P(a | \neg R)} \cdot \frac{P(b | R)}{P(b | \neg R)}$$

Generalising this result to the case of n clues and taking the “log odds” we obtain:

$$\text{LogO}(R | a_1, \dots, a_n) = \text{LogO}(R) + \sum_{i=1}^n (\text{LogO}(R | a_i) - \text{LogO}(R))$$

This is used at retrieval time to evaluate the log odds of relevance for each document in the collection with respect to the query.

- (2) A logistic regression analysis on a learning sample is used to obtain an estimate of the terms on the right hand side of the previous equation. Unfortunately, the required learning sample is often only available within the environment of test collections, although it could be possible to use the results of previous good queries for this purpose.

The estimation of $\text{LogO}(R)$ is quite straightforward using simple proportions. A more complex matter is the estimation of $\text{LogO}(R | a_i)$, when there are too few query-document pairs in the learning set with the clue a_i to yield estimates of $P(R | a_i)$ and $P(\neg R | a_i)$. To go beyond simple averaging, Cooper uses multiple logistic regression analysis. If we assume that the clue a_i is a composite clue, whose elementary attributes are h_1, \dots, h_m then we can estimate $\text{LogO}(R | a_i)$ as follows:

$$\begin{aligned} \text{LogO}(R | a_i) &= \text{LogO}(R | h_1, \dots, h_m) \\ &= c_0 + c_1 h_1 + \dots + c_m h_m \end{aligned}$$

To demonstrate how the logistic function comes into the model, the probability of relevance of a document can be expressed as:

$$P(R | h_1, \dots, h_m) = \frac{e^{c_0 + c_1 h_1 + \dots + c_m h_m}}{1 + e^{c_0 + c_1 h_1 + \dots + c_m h_m}}$$

Taking the log odds of both sides conveniently reduces this formula to the previous one.

⁶A simple clue could be, for example, the presence of an index term in a document. Clues need to be machine-detectable

- (3) A second logistic regression analysis, based on the same learning sample, is used to obtain another predictive rule for combining the composite clues and for correcting biases introduced by the simplifying assumption.

The linked dependence assumption tends to inflate the estimates for documents near the top of the output ranking whenever the clues on which the estimates are based are strongly interdependent. To help correct this, a second level logistic regression analysis is performed on the results of the first. It has the following form:

$$\text{LogO}(R | a_1, \dots, a_n) = d_0 + d_1 Z + d_2 n$$

where $Z = \sum_{i=1}^n (\text{LogO}(R | a_i) - \text{LogO}(R))$ and n is the number of composite clues. More elaborate correcting equation might also be considered.

When a query is submitted to the system and a document is compared against it the technique in part 2 is applied to evaluate the log odds necessary to obtain Z . That is then employed in part 3 to adjust estimate of the log odds of relevance for the document.

This approach seems flexible enough to handle almost any type of probabilistic retrieval clues likely to be of interest, and is especially appropriate when the retrieval clues are grouped or composite. However, the effectiveness of the methodology remains to be determined empirically, and its performance compared with other retrieval methods. An experimental investigation is currently under way by Cooper, and the use of logistic regression has also been investigated by Fuhr, as reported in the proceedings of the TREC-1 Conference [Fuhr and Buckley 1993].

3.8 The N-Poisson indexing model

This probabilistic indexing model is an extension to n-dimensions of the *2-Poisson model* proposed by Bookstein et al. in 1974 [Bookstein and Swanson 1974]. In its *2-dimensional* form the model is based upon the following assumption. If the number of occurrences of a term within a document is different depending upon whether the document is relevant or not, and if the number of occurrences of that term can be modelled using a known distribution, then it is possible to decide if an term should be assigned to a document by determining which of the two distributions the term belongs to. The 2-Poisson model resulted from a search for the statistical distribution of occurrence of potential index terms in a collection of documents.

We can extend the above idea to the *n-dimensional* case. We suppose there are n classes of documents in which the term x_i appears with different frequencies according to the extents of coverage of the topic related to that specific term. The distribution of the term within each class is governed by a single Poisson. Given a term x_i , and a document class for that term K_{ij} , and the expectation of the number of occurrences of that term in that class λ_{ij} , then the probability that a document contains l occurrence of x_i , i.e. that $tf(x_i) = l$, given that it belongs to the class K_{ij} , is given by:

$$P(tf(x_i) = l | \vec{x} \in K_{ij}) = \frac{\lambda_{ij}^l}{l!} e^{-\lambda_{ij}}$$

Extending this result, the distribution of a certain term within the whole collection of documents is governed by a sum of Poisson distributions, one for each class

of coverage. In other words, if we take a document at random in the collection, whose probability of belonging to class K_{ij} is p_{ij} then the probability of having l occurrences of term x_i is:

$$P(tf(x_i) = l) = \sum_{j=1}^n p_{ij} e^{-\lambda_{ij}} \frac{\lambda_{ij}^l}{l!}$$

This result can be used with a Bayesian inversion to evaluate $P(\vec{x} \in K_{ij} \mid tf(x_i = l))$ for retrieval purposes. The parameters λ_{ij} and p_{ij} can be estimated without feedback information by applying statistical techniques to the document collection.

Experiments have shown that the performance of this model is not always consistent. Some experiments performed by Harter [Harter 1975] on a 2-Poisson showed that a significant number of “good” index terms were 2-Poisson, but they did not provide conclusive evidence of the validity of the n-Poisson model. These results were co-validated by Robertson et al. [Robertson and Walker 1994]. They demonstrated considerable performance improvements by using some effective approximations to the 2-Poisson model on the TREC collection. Other research investigated the possibility of using a 3-Poisson, and lastly Margulis [Margulis 1992; Margulis 1993] investigated the generalised n-Poisson model on several large full text document collections. His findings were more encouraging than those of the previous work. He determined that over 70% of frequently occurring words were indeed distributed according to a n-Poisson distribution. Further, he found that the distribution of most n-Poisson words had relatively few single Poisson components, instead, usually two, three, or four. He concluded suggesting that his study provides strong evidence that the n-Poisson distribution could be used as a basis for accurate statistical modelling of large document collections. However, to date, the n-Poisson approach lacks work on retrieval strategies based upon the results gained so far.

4. UNCERTAIN INFERENCE MODELS

The models presented in this section are based on the idea that IR is a process of uncertain inference. Uncertain inference models are based on more complex forms of relevance than those used in relevance models, which are based mainly upon statistical estimations of the probability of relevance. With uncertain inference models, information not present in the query formulation may be included in the evaluation of the relevance of a document. Such information might be domain knowledge, knowledge about the user, user’s relevance feedback, etc. The estimation of the probabilities $P(R \mid q_k, d_i, K)$ involves the representation of the knowledge K .

Another characteristic of uncertain inference models is that they are not as strongly collection-dependent as relevance models. Parameters in relevance models are only valid for the current collection, while inference models can use knowledge of the user or the application domain that can be useful with many other collections.

This research area is promising in that it is attempting to move away from the traditional approaches, and may provide the breakthrough that appears necessary to overcome the limitations of current IR systems.

There are two main types of uncertain inference models. The first is based on non-classical logic, to which probabilities are mapped (section 4.1), and the second

is based on Bayesian inferences (section 4.2).

4.1 A non-classical logic for IR

In 1986, Van Rijsbergen proposed a paradigm for probabilistic IR in which IR was regarded as a process of uncertain inference [van Rijsbergen 1986]. The paradigm is based on the assumption that queries and documents can be regarded as logical formulae, and to answer a query, an IR system must prove the query from the documents. This means that a document is relevant to a query only if it implies the query; in other words, if the logical formula $d \rightarrow q$ can be proven to *hold*. The proof may use additional knowledge K ; in that case, the logical formula is then rewritten as $(d, K) \rightarrow q$.

The introduction of uncertainty comes from the consideration that a collection of documents cannot be considered as a consistent and a complete set of statements. In fact, documents in the collection could contradict each other in any particular logic, and not all the necessary knowledge is available. It has been shown [van Rijsbergen 1986; Lalmas 1997] that classical logic, the most commonly used logic, is not adequate to represent query and documents because of the intrinsic uncertainty present in IR⁷. Therefore, Van Rijsbergen proposes the *logical uncertainty principle* [van Rijsbergen 1986]:

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 principle says nothing about how “uncertainty” and “minimal” might be quantified. However, in his paper, Van Rijsbergen suggested an information-theoretic approach. This idea has been followed by Nie et al. [Nie et al. 1996] and Lalmas [van Rijsbergen and Lalmas 1996]. However, that work is somewhat beyond the scope of this paper.

More close to this survey, Van Rijsbergen [van Rijsbergen 1989] later proposed to estimate $P(d \rightarrow q)$ by *imaging*. Imaging formulates probabilities based on a “possible worlds” semantics [Stalnaker 1981]. According to this semantics, a document is represented by a possible world w ; i.e. a set of propositions with associated truth values. Let τ denote a logical truth function, then $\tau(w, p)$ denotes the truth of the proposition p in the world w . Further, let $\sigma(w, p)$ denote the world most similar to w where p is true. Then, $y \rightarrow x$ is true at w if and only if x is true at $\sigma(w, p)$.

Imaging uses this notion of most similar worlds to estimate $P(y \rightarrow x)$. Every possible world w has a probability $P(w)$, and the sum over all possible world is 1. $P(y \rightarrow x)$ is computed in the following way:

$$\begin{aligned} P(y \rightarrow x) &= \sum_w P(w) \tau(w, y \rightarrow x) \\ &= \sum_w P(w) \tau(\sigma(w, y), y \rightarrow x) \\ &= \sum_w P(w) \tau(\sigma(w, y), x) \end{aligned}$$

⁷There are other reasons why classical logic is not adequate, but these are not relevant to this paper (but see [Lalmas 1997]).

It remains undetermined how to evaluate the function σ on document representations, and further, how to assign a probability P to them. There have been a few attempts at using imaging in IR (for example [Amati and Kerpedjiev 1992; Sembok and van Rijsbergen 1993]) with rather disappointing results. A recent attempt by Crestani et al. [Crestani and van Rijsbergen 1995] taking the view that “an index term is a world” obtain better results.

In the above framework, the concept of relevance, does not feature. In [van Rijsbergen 1992] Van Rijsbergen proposed to evaluate the probability of relevance $P(R | q_k, d_i)$ using *Jeffrey’s conditionalisation*. This conditionalisation, described as “Neo-Bayesianism” by Pearl [Pearl 1990], allows conditioning to be based on evidence derived from the “passage of experience”, where the evidence can be non-propositional in nature. A comprehensive treatise of Jeffrey’s studies on probability kinematics, i.e. on how to revise a probability measure in the light of uncertain evidence or observation, can be found in [Jeffrey 1965]. By means of the famous example of inspecting the colour of a piece of cloth by candlelight in that book, Van Rijsbergen introduced a form of conditioning that has many advantages over Bayesian conditioning. In particular, it enables conditioning on uncertain evidence, and allows order-independent partial assertion of evidence. Such advantages, despite some strong assumptions, convinced van Rijsbergen that this particular form of conditionalisation is more appropriate for IR than Bayesian conditionalisation. However, despite the appeal of Jeffrey’s conditionalisation, the evaluation of the probability of relevance involves parameters, the estimation of which remain problematic.

In the same paper [van Rijsbergen 1992], Van Rijsbergen makes the connection between Jeffrey’s conditionalisation and the Dempster-Shafer’s Theory of Evidence [Dempster 1968; Shafer 1976]. This theory can be viewed as a generalisation of the Bayesian method (for example, it rejects the additivity rule), and have been used by some researchers to develop IR models (see [Schoken and Hummel 1993; de Silva and Milidiu 1993]).

4.2 The Inference Network model

When IR is regarded as a process of uncertain inference, then the calculation of the probability of relevance, and the general notion of relevance itself, becomes more complex. Relevance becomes related to the inferential process by which we find and evaluate a relation between a document and a query.

A probabilistic formalism for describing inference relations with uncertainty is provided by *Bayesian inference networks*, which have been described extensively in [Pearl 1988] and [Neapolitan 1990]. Turtle and Croft [Turtle 1990; Turtle and Croft 1990; Turtle and Croft 1991] applied such networks to IR. Figure 2 depicts an example of such a network. Nodes represent IR entities such as documents, index terms, concepts, queries, and information needs. We can choose the number and kind of nodes we wish to use according to how complex we want the representation of the document collection or the information needs to be. Arcs represent probabilistic dependencies between entities. They represent conditional probabilities; that is, the probability of an entity being true given the probabilities of its parents being true.

The inference network is usually made up of two component networks: a document network and a query network. The document network represents the docu-

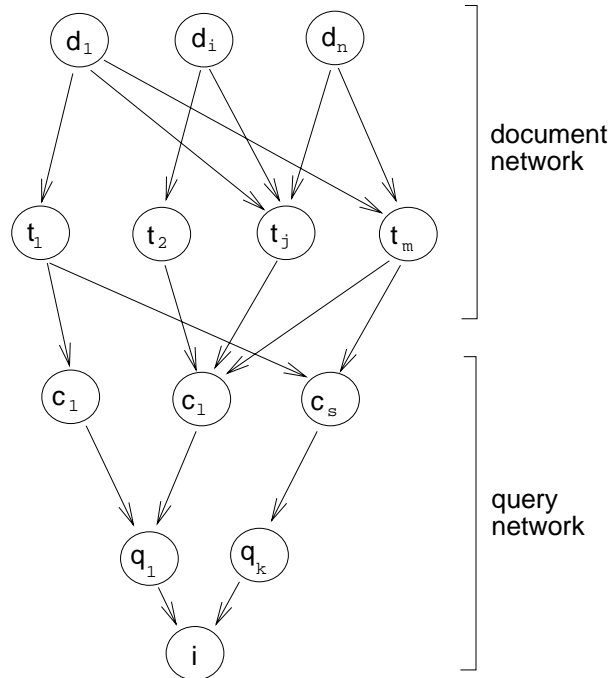


Fig. 2. An inference network for IR.

ment collection. It is built once for a given collection and its structure does not change. A query network is built for each information need and can be modified and extended during each session by the user in a interactive and dynamic way. The query network is attached to the static document network in order to process a query.

In a Bayesian inference network, the truth value of a node depends only upon the truth values of its parents. To evaluate the strength of an inference chain going from one document to the query we set the document node d_i to “true” and evaluate $P(q_k = true | d_i = true)$. This gives us an estimate of $P(d_i \rightarrow q_k)$.

It is possible to implement various traditional IR models on this network by introducing nodes representing Boolean operators or by setting appropriate conditional probability evaluation functions within nodes.

One particular characteristic of this model that warrants exploration is that multiple document and query representations can be used within the context of a particular document collection (e.g., a Boolean expression or a vector). Moreover, given a single information need, it is possible to combine results from multiple queries and from multiple search strategies.

The strength of this model comes from the fact that most classical retrieval models can be expressed in terms of a Bayesian inference network by estimating in different ways the weights in the inference network [Turtle and Croft 1992a]. Nevertheless, the characteristics of the Bayesian inference process itself, given that nodes (evidence) can only be binary (either present or not) limits its use to where

“certain evidence” [Neapolitan 1990] is available. The approach followed by van Rijsbergen (section 4.1), which makes use of “uncertain evidence” by using Jeffrey’s conditionalisation, therefore appears attractive.

5. EFFECTIVE RESULTS FROM FAULTY MODELS

Most of the probabilistic models presented in this paper use simplifying assumptions to reduce the complexity related to the application of mathematical models to real situations. There are general risks inherent in the use of such assumptions. One such risk is that there may be *inconsistencies* between the assumptions laid down and the data to which they are applied.

Another is that there may be a *misidentification of the underlying assumptions*, i.e. the stated assumptions may not be the real assumptions upon which the derived model or resulting experiments are actually based. This risk was identified by Cooper [Cooper 1995]. He identified the three most commonly adopted simplifying assumptions which are related to the statistical independence of documents, index terms, and information needs:

Absolute Independence

$$P(a, b) = P(a) \cdot P(b)$$

Conditional Independence

$$P(a, b | R) = P(a | R) \cdot P(b | R)$$

$$P(a, b | \neg R) = P(a | \neg R) \cdot P(b | \neg R)$$

These assumptions are interpreted differently whether a and b are regarded as properties of documents or of users.

Cooper pointed out how the combined use of the Absolute Independence assumption *and* either of the Conditional Independence assumptions yields logical inconsistencies. The combined use of these assumptions leads to the conclusion that $P(a, b, R) > P(a, b)$, which is contrary to the elementary laws of probability theory. Nevertheless, in most cases where these inconsistencies appeared, the faulty model used as the basis for experimental work has proved, on the whole, to be successful. Examples of this are given in [Robertson and Sparck Jones 1976] and [Fuhr and Buckley 1991].

The conclusion drawn by Cooper is that the experiments performed were actually based on somewhat different assumptions, which were, in fact, consistent. In some cases where the Absolute Independence assumption was used together with a Conditional Independence assumption, it seems that the required probability rankings could have been achieved on the basis of the Conditional Independence assumption alone. This is true of the model proposed by Maron et al. in [Maron and Kuhns 1960]. In other cases, the Conditional Independence assumptions could be replaced by the single *linked dependence* assumption:

$$\frac{P(a, b | R)}{P(a, b | \neg R)} = \frac{P(a | R)}{P(a | \neg R)} \cdot \frac{P(b | R)}{P(b | \neg R)}$$

This is a considerably weaker assumption, and it is consistent with the Absolute Independence assumption. This is true of the SLR model presented in section 3.7,

and of the BIR model (whose name seems to lose appropriateness in the light of these results) presented in section 3.2.

6. FURTHER RESEARCH

In the late nineties, we have come to realise that there is a leap to be made towards a new generation of IR systems; towards systems able to cope with increasingly demanding users, whose requirements and expectations continue to outstrip the progress being made in computing, storage, and transport technology. Faster machines and better interconnectivity enable access to enormous amounts of information. This information is not only increasing in amount, but also in complexity; for example, structured hypertexts consisting of multiple media are becoming the norm. Until recently, research in Information Retrieval has been confined to the academic world. Things are changing slowly. The success of the TREC initiative over the last four years (from [Harman 1993] to [Harman 1996]), particularly in terms of the interest shown by commercial organisations, demonstrates that there is a wider desire to produce sophisticated IR systems. The Web searching engines, which have a high profile in the wider community, increasingly utilise probabilistic techniques. It can only be hoped that this increasing awareness and interest will stimulate new research.

The requirements of the next generation of IR systems include:

Multimedia documents. The problem with multimedia document collections lies in the representation of the non-textual parts of documents, e.g. sounds, images, animations. Several approaches have been tried so far: They can be exemplified in the particular approach of attaching textual descriptions to non-textual parts, and the derivation of such descriptions by means of an inference process (e.g. [Dunlop 1991]). Nevertheless, such techniques avoid the real issue of directly handling the media. This applies not only to probabilistic models, but to all IR models.

Interactive retrieval. Current IR systems, even those providing forms of relevance feedback for the user, are still based upon the traditional iterative batch retrieval approach. Even relevance feedback acts upon a previous retrieval run to improve the quality of the successive run [Harman 1992b; Harman 1992a]. We need real interactive systems, enabling a greater variety of interaction with the user than merely query formulation and relevance feedback [Croft 1987]. User profile information, analysis of browsing actions, or user modification of probabilistic weights, for example, could all be taken into consideration [Croft and Thompson 1987; Croft et al. 1988; Croft et al. 1989; Thompson 1989; Thompson 1990a; Thompson 1990b]. The subjective, contextual, and dynamic nature of relevance is now being recognised and incorporated into probabilistic models [Campbell and van Rijsbergen 1996].

Integrated text and fact retrieval. There has been a steady development of the kinds of information being collected and stored in databases; notably, of text (unformatted data), and of 'facts' (formatted, often numerical, data). Demand is growing for the availability of systems capable of dealing with all types of data in a consistent and unified manner [Fuhr 1992a; Croft et al. 1992; Harper and Walker 1992; Fuhr 1993].

Imprecise data. The use of probabilistic modelling in IR is not only important for representing the document information content, but also for representing and

dealing with vagueness and imprecision in the query formulation and with imprecision and errors in the textual documents themselves [Fuhr 1990; Turtle and Croft 1992b]. For example, the increasing use of scanners and OCR in transferring documents from paper to electronic form, inevitably introduces imprecision (but see [Smith and Stanfill 1988]).

7. CONCLUSIONS

The major concepts and a number of probabilistic IR models have been described. We are aware that new models are being developed as we speak. A survey is always a bit dated. However, we believe we have covered the most important and the most investigated probabilistic models of IR.

It is not easy to draw conclusions from a survey of thirty years of research. It is safe to conclude that good results have been achieved but more research is required since there is considerable room for improvement. Current generation probabilistic IR systems work quite well when compared with the Boolean systems that they are replacing. A novice user using natural language input with a current generation probabilistic IR system gets, on average, better performance than an expert user with a Boolean system on the same collection. Moreover, theoretically, the probabilistic approach to IR seems inherently suitable for the representation and processing of the uncertain and imprecise information that is typical of IR. We believe that, with development, it will be capable ultimately of providing an integrated, holistic, and theoretically consistent framework for the effective retrieval of complex information.

Acknowledgements

We would like to thank the anonymous reviewers for their interesting and helpful comments.

REFERENCES

- AMATI, G. AND KERPEDIJEV, S. 1992. An Information Retrieval logical model: implementation and experiments. Technical Report Rel 5B04892 (March), Fondazione Ugo Bordoni, Roma, Italy.
- AMATI, G. AND VAN RIJSBERGEN, C. J. 1995. Probability, information and information retrieval. In *Proceedings of the First International Workshop on Information Retrieval, Uncertainty and Logic* (Glasgow, Scotland, UK, Sept. 1995).
- BIEBRICHER, P., FUHR, N., KNORZ, G., LUSTIG, G., AND SCHWANTNER, M. 1988. The automatic indexing system AIX/PHYS - from research to application. In *Proceedings of ACM SIGIR* (Grenoble, France, 1988), pp. 333-342.
- BOOKSTEIN, A. AND COOPER, W. S. 1976. A general mathematical model for information retrieval systems. *The Library Quarterly* 46, 2.
- BOOKSTEIN, A. AND SWANSON, D. 1974. Probabilistic models for automatic indexing. *Journal of the American Society for Information Science* 25, 5, 312-318.
- BORGOGNA, G. AND PASI, G. 1993. A fuzzy linguistic approach generalizing boolean information retrieval: a model and its evaluation. *Journal of the American Society for Information Science* 2, 70-82, 44.
- BRUZA, P. D. 1993. *Stratified Information Disclosure: a synthesis between Hypermedia and Information Retrieval*. Ph. D. thesis, Katholieke Universiteit Nijmegen, The Netherlands.
- BRUZA, P. D. AND VAN DER WEIDE, T. P. 1992. Stratified hypermedia structures for information disclosure. *The Computer Journal* 35, 3, 208-220.

- CAMPBELL, I. AND VAN RIJSBERGEN, C. J. 1996. The ostensive model of developing information needs. In *Proceedings of CoLIS 2* (Copenhagen, Denmark, Oct. 1996), pp. 251–268.
- CHIARAMELLA, Y. AND CHEVALLET, J. P. 1992. About retrieval models and logic. *The Computer Journal* 35, 3, 233–242.
- COOPER, W. S. 1971. A definition of relevance for information retrieval. *Information Storage and Retrieval* 7, 19–37.
- COOPER, W. S. 1995. Some inconsistencies and misnomers in probabilistic information retrieval. *ACM Transactions on Information Systems* 13, 1, 100–111.
- COOPER, W. S., GEY, F. C., AND DABNEY, D. P. 1992. Probabilistic retrieval based on staged logistic regression. In *Proceedings of ACM SIGIR* (Copenhagen, Denmark, June 1992), pp. 198–210.
- COX, D. R. 1970. *Analysis of Binary Data*. Methuen, London, UK.
- CRESTANI, F. AND VAN RIJSBERGEN, C. J. 1995. Information Retrieval by Logical Imaging. *Journal of Documentation* 51, 1, 1–15.
- CRESTANI, F. AND VAN RIJSBERGEN, C. J. 1995. Probability kinematics in information retrieval. In *Proceedings of ACM SIGIR* (Seattle, WA, USA, July 1995), pp. 291–299.
- CROFT, W. B. 1987. Approaches to Intelligent Information Retrieval. *Information Processing and Management* 23, 4, 249:254.
- CROFT, W. B. AND HARPER, D. J. 1979. Using probabilistic models of document retrieval without relevance information. *Journal of Documentation* 35, 285–295.
- CROFT, W. B., LUCIA, T. J., AND COHEN, P. R. 1988. Retrieving documents by plausible inference: a preliminary study. In *Proceedings of ACM SIGIR* (Grenoble, France, June 1988).
- CROFT, W. B., LUCIA, T. J., CRIGEAN, J., AND WILLET, P. 1989. Retrieving documents by plausible inference: an experimental study. *Information Processing and Management* 25, 6, 599–614.
- CROFT, W. B., SMITH, L. A., AND TURTLE, H. R. 1992. A loosely-coupled integration of a text retrieval system and an object-oriented database system. In *Proceedings of ACM SIGIR* (Copenhagen, Denmark, June 1992), pp. 223–232.
- CROFT, W. B. AND THOMPSON, R. H. 1987. *I³R*: a new approach to the design of Document Retrieval Systems. *Journal of the American Society for Information Science* 38, 6, 389–404.
- DE SILVA, W. T. AND MILIDIU, R. L. 1993. Belief function model for information retrieval. *Journal of the American Society of Information Science* 4, 1, 10–18.
- DEMPSTER, A. P. 1968. A generalization of the Bayesian inference. *Journal of Royal Statistical Society* 30, 205–447.
- DUNLOP, M. D. 1991. *Multimedia Information Retrieval*. Ph. D. thesis, Department of Computing Science, University of Glasgow, Glasgow, UK.
- FUHR, N. 1989. Models for retrieval with probabilistic indexing. *Information Processing and Management* 25, 1, 55–72.
- FUHR, N. 1990. A probabilistic framework for vague queries and imprecise information in databases. In *Proceedings of the International Conference on Very Large Databases* (Los Altos, CA, USA, 1990), pp. 696–707. Morgan Kaufman.
- FUHR, N. 1992a. Integration of probabilistic fact and text retrieval. In *Proceedings of ACM SIGIR* (Copenhagen, Denmark, June 1992), pp. 211–222.
- FUHR, N. 1992b. Probabilistic models in Information Retrieval. *The Computer Journal* 35, 3, 243–254.
- FUHR, N. 1993. A probabilistic relational model for the integration of ir and databases. In *Proceedings of ACM SIGIR* (Pittsburgh, PA, USA, June 1993), pp. 309–317.
- FUHR, N. AND BUCKLEY, C. 1991. A probabilistic learning approach for document indexing. *ACM Transactions on Information Systems* 9, 3, 223–248.
- FUHR, N. AND BUCKLEY, C. 1993. Optimizing document indexing and search term weighting based on probabilistic models. In D. HARMAN Ed., *The First Text Retrieval Conference (TREC-1)*, Special Publication 500-207 (Gaithersburg, MD, USA, 1993), pp. 89–100. National Institute of Standards and Technology.

- FUHR, N. AND KNOWRZ, G. 1984. Retrieval test evaluation of a rule based automatic indexing (AIR/PHYS). In C. J. VAN RIJSBERGEN Ed., *Research and development in Information Retrieval*, pp. 391–408. Cambridge University Press, Cambridge, UK.
- FUHR, N. AND PFEIFER, U. 1991. Combining model-oriented and description-oriented approaches for probabilistic indexing. In *Proceedings of ACM SIGIR* (Chicago, USA, Oct. 1991), pp. 46–56.
- FUNG, R. M., CRAWFORD, S. L., APPELBAUM, L. A., AND TONG, R. M. 1990. An architecture for probabilistic concept based information retrieval. In *Proceedings of ACM SIGIR* (Bruxelles, Belgium, Sept. 1990), pp. 455–467.
- GOOD, I. J. 1950. *Probability and the Weighing of Evidence*. Charles Griffin Symand Company Limited.
- HARMAN, D. 1992a. Relevance feedback and other query modification techniques. In W. B. FRAKES AND R. BAEZA-YATES Eds., *Information Retrieval: data structures and algorithms*, Chapter 11. Englewood Cliffs, New Jersey, USA: Prentice Hall.
- HARMAN, D. 1992b. Relevance feedback revisited. In *Proceedings of ACM SIGIR* (Copenhagen, Denmark, June 1992), pp. 1–10.
- HARMAN, D. 1993. Overview of the first TREC conference. In *Proceedings of ACM SIGIR* (Pittsburgh, PA, USA, June 1993), pp. 36–47.
- HARMAN, D. 1996. Overview of the fifth text retrieval conference (TREC-5). In *Proceeding of the TREC Conference* (Gaithersburg, MD, USA, Nov. 1996).
- HARPER, D. J. AND WALKER, A. D. M. 1992. ECLAIR: an extensible class library for Information Retrieval. *The Computer Journal* 35, 3, 256–267.
- HARTER, S. P. 1975. A probabilistic approach to automatic keyword indexing: part 1. *Journal of the American Society for Information Science* 26, 4, 197–206.
- HUIBERS, T. W. C. 1996. *An Axiomatic Theory for Information Retrieval*. Ph. D. thesis, Utrecht University, The Netherlands.
- JEFFREY, R. C. 1965. *The logic of decision*. McGraw-Hill, New York, USA.
- KWOK, K. L. 1990. Experiments with a component theory of probabilistic Information Retrieval based on single terms as document components. *ACM Transactions on Information Systems* 8, 4 (Oct.), 363–386.
- LALMAS, M. 1992. A logic model of information retrieval based on situation theory. In *Proceedings of the 14th BCS Information Retrieval Colloquium* (Lancaster, UK, Dec. 1992).
- LALMAS, M. 1996. *Theories of Information and Uncertainty for the modelling of Information Retrieval: an application of Situation Theory and Dempster-Shafer's Theory of Evidence*. Ph. D. thesis, University of Glasgow.
- LALMAS, M. 1997. Models in information retrieval: Introduction and overview. *Information Processing and Management*. In print.
- MARGULIS, E. L. 1992. N-poisson document modelling. In *Proceedings of ACM SIGIR* (Copenhagen, Denmark, June 1992), pp. 177–189.
- MARGULIS, E. L. 1993. Modelling documents with multiple Poisson distributions. *Information Processing and Management* 29, 2, 215–227.
- MARON, M. E. AND KUHNS, J. L. 1960. On relevance, probabilistic indexing and retrieval. *Journal of the ACM* 7, 216–244.
- MILLER, W. L. 1971. A probabilistic search strategy for MEDLARS. *Journal of Documentation* 27, 254–266.
- MIZZARO, S. 1996. Relevance: the whole (hi)story. Technical Report UDMI/12/96/RR (Dec.), Dipartimento di Matematica e Informatica, Universita' di Udine, Italy.
- NEAPOLITAN, R. E. 1990. *Probabilistic reasoning in expert systems*. John Wiley and Son Inc. , New York, USA.
- NIE, J. Y. 1988. An outline of a general model for information retrieval. In *Proceedings of ACM SIGIR* (Grenoble, France, June 1988), pp. 495–506.
- NIE, J. Y. 1989. An Information Retrieval model based on Modal Logic. *Information Processing and Management* 25, 5, 477–491.

- NIE, J. Y. 1992. Towards a probabilistic modal logic for semantic based information retrieval. In *Proceedings of ACM SIGIR* (Copenhagen, Denmark, June 1992), pp. 140–151.
- NIE, J. Y., LEPAGE, F., AND BRISEBOIS, M. 1996. Information retrieval as counterfactual. *The Computer Journal* 38, 8, 643–657.
- PEARL, J. 1988. *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Morgan Kaufmann, San Mateo, California.
- PEARL, J. 1990. Jeffrey's rule, passage of experience and Neo-Bayesianism. In H. E. KYBURG, R. P. LUOI, AND G. N. CARLSON Eds., *Knowledge representation and defeasible reasoning*, pp. 245–265. Kluwer Academic Publisher, Dodrecht, The Netherlands.
- ROBERTSON, S. E. 1977. The probability ranking principle in IR. *Journal of Documentation* 33, 4 (Dec.), 294–304.
- ROBERTSON, S. E., MARON, M. E., AND COOPER, W. S. 1982. Probability of relevance: a unification of two competing models for document retrieval. *Information Technology: Research and Development* 1, 1–21.
- ROBERTSON, S. E. AND SPARCK JONES, K. 1976. Relevance weighting of search terms. *Journal of the American Society for Information Science* 27, 129–146.
- ROBERTSON, S. E. AND WALKER, S. 1994. Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval. In *Proceedings of ACM SIGIR* (Dublin, Ireland, June 1994), pp. 232–241.
- SALTON, G. 1968. *Automatic information organization and retrieval*. Mc Graw Hill, New York.
- SAVOY, J. 1992. Bayesian inference networks and spreading activation in hypertext systems. *Information Processing and Management* 28, 3, 389–406.
- SCHOKEN, S. S. AND HUMMEL, R. A. 1993. On the use of dempster shafer model in information indexing and retrieval applications. *International Journal of Man-Machine Studies* 39, 1–37.
- SEBASTIANI, F. 1994. A probabilistic terminological logic for modelling information retrieval. In *Proceedings of ACM SIGIR* (Dublin, Ireland, 1994), pp. 122–131.
- SEMBOK, T. M. T. AND VAN RIJSBERGEN, C. J. 1993. Imaging: a relevance feedback retrieval with nearest neighbour clusters. In *Proceedings of the BCS Colloquium in Information Retrieval* (Glasgow, UK, March 1993), pp. 91–107.
- SERACEVIC, T. 1970. The concept of "relevance" in information science: a historical review. In T. SERACEVIC Ed., *Introduction to Information Science*, Chapter 14. R. R. Bower Company, New York, USA.
- SHAFER, G. 1976. *A Mathematical Theory of Evidence*. Princeton University Press.
- SMITH, S. AND STANFILL, C. 1988. An analysis of the effects of data corruption on text retrieval performance. Technical report (Dec.), Thinking Machines Corporation, Cambridge, MA, USA.
- SPARCK JONES, K. 1981. *Information Retrieval Experiments*. Butterworth, London.
- STALNAKER, R. 1981. Probability and conditionals. In W. L. HARPER, R. STALNAKER, AND G. PEARCE Eds., *Is*, The University of Western Ontario Series in Philosophy of Science, pp. 107–128. Dordrecht, Holland: D. Riedel Publishing Company.
- THOMPSON, P. 1990a. A combination of expert opinion approach to probabilistic Information Retrieval. Part 1: the conceptual model. *Information Processing and Management* 26, 3, 371–382.
- THOMPSON, P. 1990b. A combination of expert opinion approach to probabilistic Information Retrieval. Part2: mathematical treatment of CEO model 3. *Information Processing and Management* 26, 3, 383–394.
- THOMPSON, R. H. 1989. The design and implementation of an intelligent interface for Information Retrieval. Technical report, Computer and Information Science Department, University of Massachusetts, Amherst, MA. USA.
- TURTLE, H. 1990. *Inference Networks for Document Retrieval*. Ph. D. thesis, Computer and Information Science Department, University of Massachusetts, Amherst (USA).

- TURTLE, H. R. AND CROFT, W. B. 1990. Inference networks for document Retrieval. In *Proceedings of ACM SIGIR* (Brussels, Belgium, Sept. 1990).
- TURTLE, H. R. AND CROFT, W. B. 1991. Evaluation of an inference network-based retrieval model. *ACM Transactions on Information Systems* 9, 3 (July), 187–222.
- TURTLE, H. R. AND CROFT, W. B. 1992a. A comparison of text retrieval models. *The Computer Journal* 35, 3 (June), 279–290.
- TURTLE, H. R. AND CROFT, W. B. 1992b. Uncertainty in information retrieval systems. Unpublished paper.
- VAN RIJSBERGEN, C. J. 1977. A theoretical basis for the use of co-occurrence data in Information Retrieval. *Journal of Documentation* 33, 2 (June), 106–119.
- VAN RIJSBERGEN, C. J. 1979. *Information Retrieval* (Second ed.). Butterworths, London.
- VAN RIJSBERGEN, C. J. 1986. A non-classical logic for Information Retrieval. *The Computer Journal* 29, 6, 481–485.
- VAN RIJSBERGEN, C. J. 1989. Toward a new information logic. In *Proceedings of ACM SIGIR* (Cambridge, USA, June 1989).
- VAN RIJSBERGEN, C. J. 1992. Probabilistic retrieval revisited. Departmental Research Report 1992/R2 (Jan.), Computing Science Department, University of Glasgow, Glasgow, UK.
- VAN RIJSBERGEN, C. J. AND LALMAS, M. 1996. An information calculus for information retrieval. *Journal of the American Society of Information Science* 47, 5, 385–398.
- WONG, S. K. M. AND YAO, Y. Y. 1989. A probability distribution model for Information Retrieval. *Information Processing and Management* 25, 1, 39–53.
- WONG, S. K. M. AND YAO, Y. Y. 1995. On modelling information retrieval with probabilistic inference. *ACM Transactions on Information Systems* 13, 1, 38–68.
- ZADEH, L. A. 1987. *Fuzzy sets and Applications: Selected Papers*. Wiley, New York.