# Information retrieval and artificial intelligence

Karen Sparck Jones
Computer Laboratory, University of Cambridge
New Museums Site, Pembroke Street
Cambridge CB2 3QG, England
*sparckjones@cl.cam.ac.uk*

February 1999
Appeared in *Artificial Intelligence, 114 (1-2), 1999, 257-281*

**Abstract**

This paper addresses the relations between information retrieval (IR) and AI. It examines document retrieval, summarising its essential features and illustrating the state of its art by presenting one probabilistic model in detail, with some test results showing its value. The paper then analyses this model and related successful approaches, concentrating on and justifying their use of weak, redundant representation and reasoning. It goes on to other information management tasks and considers how the concepts and methods developed for retrieval may be applied to these, concluding by arguing that the ways of dealing with information may also have wider relevance to AI.

## 1 Introduction

"Information retrieval" is an all-embracing term. In this paper it is taken in its well-established sense of *document retrieval*. I shall further restrict it, initially, to written document, or text retrieval, returning later to consider other types of document, for instance images.

This paper addresses the question: What has information retrieval (in the sense of document retrieval) to do with artificial intelligence? The answer may appear obvious, namely everything. If IR means, as it most importantly and challengingly does, automatic content-based information retrieval, then the usual assumption in AI is that AI researchers will show IR practitioners how to do this. If IR is seen as a search for unknown, and under-specified, information in a world of information as conveyed by natural-language texts, it is easy to conclude that what AI discovers about the representation of knowledge, reasoning under uncertainty, and acquiring new knowledge, will be clearly applicable to document retrieval.

The purpose of this paper is to argue that the boot, or at least a slipper, is on the other foot; that when key properties of document retrieval are recognised it should not, in particular, be conflated with the question answering that is a critical focus of AI; that information seeking as supported by document retrieval is a related activity that AI needs to be alert to; and that the principles that are applied and technologies that have been developed in the past forty years of IR research have important lessons for AI. As the late Gerard Salton noted (Salton 1995), AI researchers have persistently claimed that IR can only be done properly with the aid of AI techniques, for instance rule-based systems; but they have never delivered better performance than IR researchers have achieved with their own, apparently non-AI, methods. My argument here is that this AI failure has not been because developing AI techniques suited to content-based information retrieval is hard. The failure has arisen because AI researchers have misunderstood IR, and have not appreciated that the resilience of IR's own methods has been because they are intrinsically better suited to the task. However this does not imply that IR has nothing to do with AI: they share the *general* task of finding information. Thus my further claim is that the methods that have been developed for document retrieval have some lessons for, and potential use within, information management as an AI concern.

Well-known document retrieval methods seem surprisingly simple: with single words as terms, weighted using frequency data, and with query-document matching scores as functions of conjoined term matches, levels of performance can be reached that are as good as those obtained with much more elaborate techniques aimed at explicitly capturing elements of request and document meaning and at exploiting conceptual relations between these. These levels of performance are also absolutely quite respectable, especially given the sparse requests that users often submit.

In the next section, 2, I shall review the essential elements and issues of document retrieval: though document retrieval is a familiar human activity, some analysis is needed to explain the successful strategies that have been developed in IR research. Section 3 addresses the models on which retrieval systems are based. It presents one particular model in some detail, using some recent experimental results to illustrate its key features, and then relates it to some other leading approaches. In Section 4 I discuss these strategies from an AI point of view, characterising them in AI terms. I then consider, in Section 5, how the methods developed for 'classical' document retrieval can be extended both to different types of document, and to related information seeking and management tasks, where further connections with AI may be made. In conclusion, I summarise the lessons that IR, as exemplified by document retrieval, may have for AI.

## 2 Essentials

The central situation in information retrieval is that the inquirer, or *user*, wants to find out about something, or wants to learn more about something, i.e. the user has an information *need*. This internal need may be expressed in natural language, through the user's information *request*, e.g. `I want information about the manufacture of Navaho blankets`. This expression of need may be less or more full, e.g. `... manufacture of Navaho blankets before the introduction of commercial dyes`, and explicitly motivated, e.g. `... blankets because I am interested in American Indian crafts`, and so forth. The two critically important points, however, are first, that even if the information need is expressed linguistically as a question, e.g. `How are Navaho blankets made?`, the question is open-ended, seeking information not answers (unlike e.g. `Were Navaho blankets made before 1900 of cotton?`; and second, that the user's statement of need, whether or not it has the form of a question, is necessarily under-specific. The user is seeking information precisely because they do not know what they want to know about $X$.

Of course the richness and range of the notion of information need implies that there is no absolute distinction between information-seeking and answer-seeking; but there is nevertheless a large area where the user, even when starting with a question, is naturally interested in what elaborates, supports, explains, etc the answer, i.e. in being told about something. Similarly there is no absolute distinction between the case where the user is knowingly seeking new documents as opposed to recovering previously-seen but inadequately remembered ones (`Find those documents on Navaho weaving`).

Clearly, the more the user supplies about the context of their need, the better identified and scoped their information need is. But it is by definition incompletely characterised, implying that there is no certainty the user's information 'hole' could, given an informed system, be filled, reduced, or even located. [1] For instance, is `manufacture` to do with type of weaving frame? Does the user know what a weaving frame is? etc.

The foregoing refers to the 'substance' of the user's need. The IR situation is further distinguished by the particular form of response sought, namely documents about e.g. the manufacture of Navaho blankets. Because the user is seeking to remedy a lack of information, and can only do this effectively if the information supplied is itself coherent and motivated (broadly speaking, elaborated and explained), and can be properly connected with what they know already, it is natural to seek a whole *discourse* as a response, and also to prefer a document in which the discourse on the user's topic figures largely.

However just as the user's request is only an approximate expression of their need (because they are not yet informed), so is a document text as an expression of its conceptual content (though it is hopefully rather better because the author is already informed). Retrieval is thus dealing with two 'inaccessibles', information need and information content and, further, with the relation between the two represented by the statement that the document is *relevant* to the user's need. Relevance is determined by the user in an assessment that identifies, i.e. *constructs*, a connection between document content and need. The user's constructive interpretation of document relevance is critical. It implies that a far more powerful AI system than any we can realistically foresee will not be able to ensure that answers it could give to questions extracted from the user's request would be appropriate. Not merely does the system not know enough about the user, it cannot. Clearly, in an extended dialogue, a process of information exchange could allow the system to determine more accurately what the user was seking, but the user's state will also be changing under the supply of information. Relevance is indeed a function of the specific situation.

Furthermore, the presumption that an AI system with an integrated knowledge base would be superior to a file of documents misses the point that individual document texts have their own value as accounts of the information their authors want to convey. Thus eviscerating documents to construct knowledge bases, and in the process sanitising their language, loses the critical element of who said what, which includes *how* they said it.

Document retrieval thus has status in its own right as a form of mediation between two users of language, the information seeker and the information supplier (or rather between two information suppliers). The requirement in system design is to produce request and document characterisations, i.e. *index descriptions*, and description comparison methods, i.e. *matching functions*, that are best predictors of relevance. These descriptions and functions have to overcome the uncertainties inherent in the situation, both those following from need and content inaccessibility and those associated with the use of natural language for requests and documents, with all its variability of expression. There is also a further uncertainty in differences in information grain: compared with documents, requests are typically minimal or reductive concept characterisations, and so where they match documents this involves a

[1] Or, in Belkin's words, the user's 'anomalous state of knowledge' (Belkin, Oddy and Brooks 1982) be rectified.

reductive encapsulation of document content which introduces a new uncertainty.

Given these essential properties of the retrieval situation, it is even more surprising that the effective research methods that have been developed are as simple as they are. After presenting these in the next section, I shall return to why they are effective.

## 3 Models

Modern document retrieval systems are rooted in responses to the growth of the scientific literature, and the use of punched-card machines before the advent of computers. Traditional library systems relied on comprehensive classifications (sometimes viewed as characterisations of the world, sometimes only as characterisations of views of the world), typically allocating a document (book) to a single slot (shelf location). Though valuable in providing a search hierarchy, these schemes were perceived as too simplistic and restrictive for the growing journal and technical literature that required more specialised and also more flexible means of description. Flexibility was needed to allow for the different relational structures within which very specific concepts and topics could be embedded, e.g. a document on Navaho blankets could relate to textile entities, manufacturing processes, aesthetic styles and practical or cultural uses.

The most important development, for subsequent IR, was the idea of *descriptors* (Mooers 1951) (Taube 1955): concept labels that could be used in free combination for individual documents or requests and which could consequently allow *partial* matching on descriptor sets if complete matches could not be obtained, as a way of compensating for variation in detailed concept coverage or topic perspective, e.g. `development Navaho blanket` and `manufacture Navaho blanket` at least share `Navaho blanket`. Descriptor vocabularies, like the earlier class labels, along with simple grammar rules for modes of combining them, were in particular intended to overcome natural language variability in actual request and document texts, e.g. `MANUFACTURE` for "manufacture", "preparation". In the guise of sets of subject headings, and thesauri, these vocabularies became the staple of commercial automated retrieval systems from the sixties onwards, enhanced by the power of terminological relations like 'broader than', 'see also'.

However though vocabulary normalisation appeared a powerful aid to effective retrieval, it has a downside in the effort needed to assign headings to documents in indexing (though this is sometimes done automatically (Biebricher et al. 1988)) and to develop and maintain thesauri, and in the professional knowledge needed to form appropriate search specifications from the user's actual request. More importantly, the flexibility of topic specifications that such descriptor sets in principle allow was offset by the normal requirement for Boolean query specifications, with all or nothing matches on Boolean term conjunctions. Even where conjoints are OR-groups, the requirement to match on all AND-conjoints is too crude. Thus while the user would prefer documents that matched on the full query `(MANUFACTURE OR PREPARATION) AND NAVAHO AND BLANKET`, this approach means that the user will not get documents that just match on `NAVAHO AND BLANKET` though these may be the best available, or in practice sufficient for the user (or indeed may in fact be about manufacture but not indexed as such). Replacing variable natural language by a normalising controlled vocabulary and prescriptive basic syntax seems desirable, but in practice has been found over-restrictive, even allowing for the modest scale and simple form that indexing descriptions, as leads into documents, should have.

Modern IR theory and practice are responses to the retrieval situation that overcome the limitations of these first generation automated systems, and in addition take advantage of the much increased computing power that allows full-text searching. First, these approaches deliberately use the natural language of requests and document texts, on the basis that this is critical for content representation, applying other means of compensating for vocabulary, syntax and perspective variation. Second, they offer partial matching in an open way that includes allowing for alternative indexing, responding to the facts of the document file rather than the user's expectation or prescription; this is preferable to the common approach that uses natural rather than controlled terms but still with Boolean operators. Third, they systematically exploit properties of the way words are actually used in the file for more discriminating and sensitive matching than the conventional apparatus, which at best responds to this reality only in an adhoc and crude way through decisions about the system's terminology and inidividual Boolean request forms.

Modern approaches are statistically based, or at least statistically flavoured, in the sense of exploiting frequency data about the occurrences and cooccurrences of natural language *terms* (single words or compounds, full form or stemmed). That is, they assume a correlation between term frequency and conceptual salience. This correlation is not necessarily treated in the most obvious way: thus since retrieval is a selective business aimed at identifying the normally few relevant documents amidst the mass of non-relevant ones, the more a search term occurs in the file, the less discriminating it is.

It is clearly desirable to have a formal model as a base for defining indexing and searching functions. A probabilistic base for IR was first proposed in the sixties by Maron and Kuhns (1960), and the probabilistic model developed by Robertson and his colleagues and extensively tested via the City University's Okapi system has proved theoretically powerful and practically effective (Okapi 1997). It is summarised in the next section, to illustrate the type of model that has been found suited to document retrieval: for a much fuller account see Sparck Jones, Walker and Robertson (1998). I shall consider its relation to some other leading models later.

## 3.1 A probabilistic model

### 3.1.1 Basics

We assume *descriptions* of documents and requests, i.e. some representation of their content (intention, etc), and of *terms* as elements of descriptions. These can be of any form, variously related to both content and actual text: I will return to this point later. For the moment we assume descriptions are sets of single words, such that the set of terms for the document file constitutes the system's term vocabulary. Each term may be present in or absent from a document (or request) description, or may have associated information such as frequency of occurrence in the document. For future simplicity and following widespread convention in presenting the model we refer to document descriptions just as documents, and request descriptions as queries. The object of the model is then to replace *initial* given descriptions by *final* ones that determine the system's retrieved output for the user.

The model defines the aim of a retrieval system as answering, for each query and each document, the Basic Question:

- What is the probability that this document is relevant to this query?

where relevance is the result of a user decision motivated by their information need. The role of the Basic Question, within the model framework, is to rank the file documents in order of their probability of relevance: thus the Probability Ranking Principle (Robertson 1977) states that:

- If retrieved documents are ordered by decreasing probability of relevance on the data available, then the system's effectiveness is the best to be gotten for the data.

(This can be related to a user decision rule for inspecting the document (van Rijsbergen 1979).)

Then if we have a document $D$ and query $Q$ we have two events, $L$, that $D$ is *liked*, i.e. is relevant for $Q$, and $\overline{L}$, that $D$ is not *not liked*, i.e. is not relevant. Clearly we want to calculate the probability $P(L|D)$; however to allow for elaboration on the *attributes* of $D$, we apply Bayes' Theorem expressing $P(L|D)$ in terms of $P(D|L)$, and further, for later convenience, replace the probability by log-odds to obtain:

$$\log \frac{P(L|D)}{P(\overline{L}|D)} = \log \frac{P(D|L)}{P(D|\overline{L})} \; + \; \log \frac{P(L)}{P(\overline{L})} \tag{1}$$

Then introducing the notion of matching score, $MS$, as a function of descriptions and $MS(D)$ as the score for an individual document in relation to a particular query, we start by defining the most primitive matching score:

$$\mathit{MS\text{-}PRIM}(D) = \log \frac{P(D|L)}{P(D|\overline{L})} \tag{2}$$

i.e.

$$\mathit{MS\text{-}PRIM}(D) = \log \frac{P(L|D)}{P(\overline{L}|D)} \; - \; \log \frac{P(L)}{P(\overline{L})} \tag{3}$$

The model makes the strong Independence Assumption about document attributes, namely:

- Given relevance (likedness), the attributes are statistically independent.

In reality attributes like natural language words are not independent. But the Assumption makes model development easier, and in practice has allowed easier model implementation without significant loss of performance, mainly because the dependence information that is available is too weak for the coarse requirements of the retrieval task. (A version of the model with less strong assumptions is given in (Cooper 1995).) Now, using the Assumption, we can define a document's probability as liked from its attribute probabilities, as

$$P(D|L) = \prod_i P(A_i = a_i|L) \tag{4}$$

where $A_i$ is the $i$th attribute, and $a_i$ is its value for the specific document, and analogously for not liked; and we then reformulate:

$$\mathit{MS\text{-}PRIM}(D) = \sum_i \log \frac{P(A_i = a_i|L)}{P(A_i = a_i|\overline{L})} \tag{5}$$

This presumes a contribution from every attribute, whether present in a document or absent from it. But given that we are primarily interested in what documents share with queries, and we are also interested in what terms the query has rather than those it does not have, we can most usefully work with a matching score based only on shared term presence:

$$MS\text{-}BASIC(D) = \sum_i \log \frac{P(A_i = a_i|L)P(A_i = 0|\overline{L})}{P(A_i = a_i|\overline{L})P(A_i = 0|L)} \tag{6}$$

(*MS-BASIC* shifts the zero of *MS-PRIM*, so that a document with none of the attributes has zero score.) Then if we define

$$W(A_i = a_i) = \log \frac{P(A_i = a_i|L)P(A_i = 0|\overline{L})}{P(A_i = a_i|\overline{L})P(A_i = 0|L)} \tag{7}$$

this gives a *weight* for the value of each attribute, so the matching score for a document is simply the sume of weights, i.e.

$$MS\text{-}BASIC = \sum_i W(A_i = a_i) \tag{8}$$

Now, turning to what attributes may be, if we define the attribute $A_i$ simply as the presence or absence of a term $t_i$, and denote $P(t_i \text{ present}|L)$ by $p_i$ and $P(t_i \text{ present}|\overline{L})$ by $\overline{p_i}$, then the formula for $W$ (equation 7) gives a weight for term presence

$$w_i = \log \frac{p_i(1 - \overline{p_i})}{\overline{p_i}(1 - p_i)} \tag{9}$$

where the matching score for a document is just the sum of the weights of the present terms.

If we have no more sophisticated definition of attribute than this, we obtain the simplest possible approach to retrieval using term coordination, where the score for a document is simply the number of terms shared with the request (so it is also labelled *unweighted (UW)* matching). This clearly supplies an intuitive ranking with documents sharing most terms with a query at highest rank, in a way that allows for partial matches and alternative ones with the same number of shared terms. This outcome would not in itself require such elaborate model development. But it should be emphasised that in the classical Cranfield experiments conducted in the sixties (Cleverdon 1967), comparing different indexing languages, simple extracted natural-language index terms applied on this basis performed as well as terms from carefully controlled index languages (albeit in the favourable circumstances of fairly specific requests and short, manually formed document descriptions).

### 3.1.2 Development

The development and validation of the model since the sixties has been by incorporating more information, i.e. by the elaboration of document (and request) attributes. These exploit on the one hand more facts about the distribution of terms across the file, and on the other facts about distributions within individual documents.

For the former, pursuing the implications of the desire to select the few relevant documents from the many non-relevant in the file leads to a term weighting function that is inversely related to frequency, namely

$$CFW = \log \frac{N}{n_i} \tag{10}$$

where $N$ is the size (number of documents in) the collection and $n_i$ is the number of documents containing query term $i$, so the matching score (equation 8) becomes

$$\textit{MS-CFW} = \sum_i \log \frac{N}{n_i} \tag{11}$$

summed over query terms. (Collection frequency is also called Inverse Document Frequency, $IDF$.)

Clearly, file-based weighting can be refined when information is available about subfiles, and in particular that some subset of the file is relevant to the query, as could occur in iterative searching. Then where $R$ is the nunber of known relevant documents for the query and $r_i$ the number of relevant in which term $i$ occurs, we can derive the new term presence weighting function 9 as (omitting the index $i$ for simplicity)

$$w = \log \frac{r(N-n-R+r)}{(R-r)(n-r)} \tag{12}$$

The earlier $CFW$ is the impoverished version of this case.

The function just given assumes complete relevance information. This is not normally available - indeed if it was there would be no point in continuing searching - and the known relevance set may in fact be very small. Our estimates of $p$ and $\overline{p}$ have therefore to allow for uncertainty. This is simply achieved by adding 0.5, to obtain the relevance weighting formula

$$RW = \log \frac{(r+0.5)(N-n-R+r+0.5)}{(R-r+0.5)(n-r+0.5)} \tag{13}$$

with associated matching score

$$\textit{MS-RW} = \sum_i \log \frac{(r_i+0.5)(N-n_i-R+r_i+0.5)}{(R-r_i+0.5)(n_i-r_i+0.5)} \tag{14}$$

This clearly favours terms with a low file incidence and high relevance incidence, which is what we want. Applying the formula retrospectively, without 0.5, usefully defines a performance upper bound. It can also be applied via imputed relevance, in so-called *blind feedback*, where top-ranking documents from an initial scan of the file are deemed to be relevant and used accordingly.

The foregoing relates query term value to one aspect of document individuality. The other important factor is the frequency with which a terms occurs within a document. This can be treated, following Harter (1975), by exploiting Poisson distributions.

The essential notion (see Sparck Jones, Walker and Robertson (1998)) is that of *eliteness*, the extent to which a document is about the topic represented by the term, which is in turn taken as represented by $TF_i$ (or $TF$ for short), the frequency of the term in the document. (For the moment assume documents are all the same length.) We can relate term frequency to the probability of a document being liked:

$$P(TF|L) = P(TF|E)P(E|L) + P(TF|\overline{E})P(\overline{E}|L) \tag{15}$$

and we can further (cf. Robertson and Walker (1994)) interpret the $A_i = a_i$ of formulae 7 and 8 as $A_i = TF_i$ or just as $TF_i$, so $W(TF_i)$ becomes a function of the term's Poisson distribution and document eliteness. This in principle gives a complex term weighting function involving inaccessible notions (eliteness) and hairy estimation; but there is an effective, much simpler approximation, where

$$W(TF_i) = \frac{TF_i(k_1 + 1)}{k_1 + TF_i}\, w_i \tag{16}$$

(The constant $k_1$ determines the weight's reaction to increasing $TF$, and can be treated as a collection-specific tuning constant; it appears to be highly non-linear with a small effective range.)

But we need more than $TF$ alone: $TF$ has clearly to be related to document length, since larger $TF$ simply because a document is longer does not imply that a document is more relevant to a user's need. Indeed it is reasonable to assume that (with respect to relevance), greater document length is just correlated with greater wordiness. Allowing for document length can be appropriately done using a normalising factor $NF = ((1 - b) + b\frac{DL}{AVDL})$ where $b$ is another constant that can accommodate levels of wordiness. This leads to the weighting formula

$$W(TF_i) = \frac{TF_i(k_1 + 1)}{k_1 * ((1 - b) + b\frac{DL}{AVDL}) + TF_i}\, w_i \tag{17}$$

which can be usefully simplified to

$$W(TF_i) = \frac{TF_i(k_1 + 1)}{K + TF_i}\, w_i \tag{18}$$

where $K = k_1 * ((1 - b) + b\frac{DL}{AVDL})$.

Then, when we take into account the two levels of file information, i.e. without and with relevance data, that we considered earlier, we can combine these with the $TF$ element to obtain further, more comprehensive term weighting functions. First, the simpler combined weight without relevance information:

$$CW = \frac{TF(k_1 + 1)}{K + TF}\ \log \frac{N}{n} \tag{19}$$

with corresponding matching score *MS-CW* by summing weights; and second the combined iterative weight, assuming at least one search iteration to gather some relevance information:

$$CIW = \frac{TF(k_1 + 1)}{K + TF}\ \log \frac{(r + 0.5)(N - n - R + r + 0.5)}{(R - r + 0.5)(n - r + 0.5)} \tag{20}$$

with matching score *MS-CIW*.

Finally, while we have so far allowed only for terms occurring more than once in documents, we could clearly have long requests (perhaps sample documents) with the same property. However in this case all that is required is to multiply weights as just defined by $QTF$, giving

$$QACW = CW\ *\ QTF \tag{21}$$

and

$$QACIW = CIW\ *\ QTF \tag{22}$$

and corresponding matching scores *MS-QACW* and *MS-QACIW* respectively.

This model-based approach has been shown to be effective in a long series of experiments, and in particular through the City University's performance in the DARPA/NIST Text REtrieval Conferences (TRECs) from 1992 onwards (TREC 1992-1998) (Sparck Jones 1998b). These tests demonstrate the importance of robust, well-motivated term weighting schemes for full text. Thus simple unweighted term matching, defining *baseline* performance, is abysmal compared with the use of $CW$, which can be taken as a *benchmark*. Some illustrative performance comparisons will be given later.

#### 3.1.3 Elaboration

The model as outlined so far provides a basic way of treating the given simple term descriptions for documents and requests, taking some now-standard preliminaries for granted: applying a *stoplist* - usually of function words - and *stemming* to reduce word form variation. There are two important elaborations of this basic scheme: these are the use of *query expansion*, and that of subdocuments or *passages*.

Query expansion is a natural idea when, as is often in practice the case, the user's initial requests are very brief, regardless of whether the initial request terms are very specific (e.g. proper names) or not. Thus enlarging requests allows both for more discriminating retrieval through matches on several terms and for more file coverage through getting a match at all. Developing this idea by exploiting statistical associations between terms in the file has a long pedigree but, somewhat surprisingly, little demonstrated value. In fact value for expansion comes only when it is controlled by reference to relevance information, with the same relative effect on performance as when $RW$ is compared with $CFW$. Thus while one may in the abstract take the initial request as only one possible rendering of the user's need and consider *all* the terms in the vocabulary as candidates for need expression, it is more reasonable to assume that the given request terms do have some superior status, simply because they are what the user has offered, and then consider other terms that are linked to them in an especially strong way through their presence in matching relevant documents.

This idea, initially developed by Rocchio (1965) and Ide (1968), is modelled probabilistically in Robertson (1990), leading to a simple formula defining the *selection value* for additional query terms. Thus given some retrieved documents and relevance assessments for these, the terms in the known-relevant subfile are ranked by offer weight:

$$OW = r RW \tag{23}$$

(ignoring subscripts). Unfortunately there is no known reliable way of determining the precise degree of expansion automatically, so it is either by user decision or an empirically-determined default. Selected terms are weighted in the usual way for the new search.

Relevance expansion, even with only a few known relevant documents, is effective, and has encouraged the use of *blind expansion* though this is only modestly useful.

With full text retrieval, especially when documents are long, it is also natural to consider passage constraints on retrieval strategies: for instance preferring texts where query term matches are close together to those where they are scattered, or taking expansion terms only from the neighbourhoods of query term matches. In principle the model applies to both of these; but as with expansion, it is necessary to find practically viable ways of determining appropriate passage sizes and experiments with passages within the model framework have so far been relatively limited and unproductive.

Finally, the model assumes term units, and these have so far been taken as simple single word stems. However given that retrieval deals in text, it is natural to consider the use of *compound terms* or *phrases* (Sparck Jones 1998a). The presumption is that proper topic characterisation implies more refined descriptors that capture more specialised relations between elements than those defined just by conjoint simple term matching (e.g. we need to recognise the fact that in "Navaho blankets" we are dealing with a sort of blanket and not a sort of Navaho); and controlled index languages frequently have structured compound terms.

The probabilistic model just summarised has no means of capturing such syntactic dependencies. But since the Independence Assumption is manifestly an oversimplification, attempts have been made to develop the model to allow for term dependencies, even if these are explicitly only associative and not syntactic. However as with other intuitively plausible elaborations of the simple model, implementation is hairy and in practice there is typically not rich enough data to distinguish specific, useful associations. Query expansion is an indirect way of picking up term relationships, but one which is more effective through being constrained by relevance information.

The model is straightforwardly applicable to the use of independently-defined compound terms, treating them as units like any others along with their member terms. But extensive research by e.g. Fagan (1989) and Mitra et al. (1997) has disconcertingly shown that complex terms defined in a linguistically sophisticated way are no better than crudely-defined ones; and indeed the general finding is that compound terms do not contribute much to retrieval performance, certainly not as much as those engaged with natural language processing (NLP) would expect.

### 3.1.4 Performance illustrations

The claim in the foregoing is that the probabilistic model provides a natural way of responding to readily available data, and key realities, of the retrieval situation. Experiments with it show that it is an appropriate, because effective, model; so they also throw light on more and less important factors affecting performance.

These experiments have been carried out over a long period with very different data sets, most recently on a large scale with full-text documents within the TREC evaluation programme (see Okapi (1997), City University papers in TREC (1992-1998), and Sparck Jones (1999)). In the comparisons between teams exploiting a whole range of strategies shown in (Sparck Jones 1998c), City University's Okapi system applying the model figures among the better-performing systems. With the exception of relevance assessment in iterative searching, indexing and searching are wholly automatic. This is important since, though really intensive manual query development, especially by skilled or informed users, can materially improve performance, many users are unwilling to make the effort to work hard on their queries. The approach also has the major advantage that it is *request-based*, since elaborate document processing is intrinsically wasteful as well as effortful.

Figure 1 shows runs selected from recent experiments designed to allow systematic comparisons across a range of retrieval devices and device combinations defined by the model. This study is fully detailed in (Sparck Jones, Walker and Robertson 1998): the results given here are intended only to indicate the typical relative contribution of the successive model elements described above to retrieval performance, as illustrated by a given data set. This *test collection* is drawn from the TREC evaluation materials and consists of some 741,000 full text documents and 150 requests, supported by the (previously obtained) relevance assessments

TREC 'T741000X' TEST COLLECTION :

741856 documents in news, computing, official publications and energy represented by full text (over 2/3) or abstracts;
150 requests, medium length.

| | no docs | no terms | av terms/ doc | no reqs | av terms/ req | no reldocs | av reldocs/ req |
|---|---|---|---|---|---|---|---|
| full set | 741856 | 1290709 | 129.9 | 150 | 10.3 | 37819 | 252.1 |
| half set | 370928 | 839463 | | | | 18927 | 126.1 |

(For further details see Sparck Jones, Walker and Robertson (1998).)

RETRIEVAL PERFORMANCE :

Using half set for searching, with relevance data drawn from the other half.
Standard performance measures, see Harman (1997).

| | AveP | P10 | P30 | P100 | P30R |
|---|---|---|---|---|---|
| UW - term coord | .036 | .110 | .093 | .071 | .048 |
| CFW - coll freq wts | .074 | .157 | .146 | .113 | .104 |
| RW - rel wts : pred rel in 10 | .139 | .223 | .227 | .182 | .201 |
| CW - comb wts | .226 | .481 | .397 | .273 | .320 |
| QACW - comb wts : query adj | .269 | .524 | .442 | .310 | .374 |
| QACIW - comb iter wts, adj : pred rel in 10 | .282 | .541 | .456 | .324 | .399 |
| QACIW + E - comb iter, adj, exp : pred rel in 10 exp 24 | .336 | .616 | .517 | .362 | .471 |

150 requests, very short length : av terms/req = 4.0

| | AveP | P10 | P30 | P100 | P30R |
|---|---|---|---|---|---|
| UW - term coord | .087 | .159 | .154 | .125 | .131 |
| QACW - comb wts : query adj | .244 | .473 | .403 | .283 | .343 |
| QACIW + E - comb iter, adj, exp : pred rel in 10 exp 16 | .260 | .516 | .433 | .295 | .360 |

Figure 1: Retrieval test data and performance illustration

that allow performance measurement. Further details of the data are given in Figure 1: note that as explained below, the document set is used in two halves, so the search set is actually about 370,000 documents.

As with all of the experimental work referred to earlier, performance is characterised using conventional measures based on *precision* and *recall*, respectively the proportion of retrieval documents that are relevant, and the proportion of relevant documents that are retrieved. The Table shows performance measured with precision at document rank cutoffs 10 and 30, *PDoc10* and *PDoc30*; precision at recall 30%, *P30R*; and average precision, *AveP*, a global measure over all relevant retrieved (for detailed definitions see Harman (1997)). These measures are related to user interests: for instance PDoc10 fits well with the output displays that World Wide Web search engines typically provide. But they are applied in a relatively abstract way: the widely-used AveP is computed over far more documents than users would normally inspect, and with large files this could also be true for P30R. In such controlled, laboratory experiments there is also no true iterative searching with users in the the cycle: tests with searching that exploit relevance data use a strategy where a document set is divided so weights computed from relevance data for one half are applied in searching the other. This technique does not allow, for instance, for modification of the user's need over a sequence of document inspections.

Evaluating interactive systems with real users, and taking account of all the other factors involved (e.g. forms of information display), is a major challenge for the field. Thus the characterisation of system performance as illustrated here has to be seen as pared down; it focuses only on systems' or devices' relative abilities to retrieve individual relevant documents in response to a static need, without regard to the complex realities of information seeking as a constructive process and with only a limited and somewhat artifical treatment of user feedback. However since obtaining relevant documents is the key requirement in retrieval, the information given by the conventional measures is valuable for real system design, especially when strategies exhibit similar relative performance over very different data sets (and hence, implicitly, user sets).

Figure 1 shows quite clearly that baseline performance is extremely poor; that as weighting become more sophisticated with the progression from *CFW* to *CW* and *QACW* performance improves; that exploiting relevance information in the corresponding *RW* versus *CIW* and *QACIW* cases is also an advantage; and that query expansion (*QACIW+E*) can be a further bonus, but only when based on real and not imputed relevance information. As discused in (Sparck Jones, Walker and Robertson 1998), these large differences are statistically significant, and the relative merits of the various devices apply across different request sets for the same documents, except for very short requests in some cases: these present a practical problem considered later as a general one for automated retrieval methods. However even here, as the Figure also shows, the major progression from *UW* to *QACW* to *QACIW+E* still gives material (and statistically significant) performance gains.

### 3.2 Other models

The probabilistic model just presented is a successful, but not the only successful, IR model. Comparing it with two other successful models thus throws useful light on IR from an AI point of view.

Salton's vector space model, implemented in the Cornell SMART system, dates from the sixties. It has provided a hospitable framework for studying a range of data types and information management tasks, undergoing continuous extension, and has proved very effective in practical implementation (see (Salton and Buckley 1988) (Salton and Buckley 1990)) and Cornell papers in TREC (1992-1998). The underlying metaphor (Salton 1975) is of an information space that can accommodate many types of entity (documents, queries, terms etc), so indexing strategies are designed to position, or reposition, documents and queries in a manner which brings queries and relevant documents close together, while pushing non-relevant documents further away. The key operations are thus vector comparison and vector modification.

The index vectors, e.g. for documents, can be defined by any attributes and can consist of differently-defined subvectors: the main constraint is the basic one of attribute conjunction for object characterisation and as the base for object comparisons. The flexibility of the vector model has the advantage of allowing very different types of attribute, e.g. citations or words as descriptors, compound or simple terms, etc; but it also has the disadvantage that it offers very little guidance as to the choice and treatment of data and task features, encouraging ad hoc rather than principled indexing and matching.

The probabilistic network model used in the INQUERY system has also been very successful in the TREC Programme evaluations, and appears to combine the advantages of a probabilistic grounding with the ability to exploit 'composite' indexing and searching devices that are outside the scope of the straightforward probabilistic or vector space models: see Turtle and Croft (1991) and University of Massachusetts papers in TREC (1992-1998). Thus it can support *fusion* methods in a a more comprehensive style than e.g. just combining different sorts of index terms. For instance it can easily support radically different entity representations, say queries as conventional Boolean expressions and as weighted term lists, and combine matching on each of these to deliver a final document score. These capabilities are an appropriate response to the fact that information is intrinsically multifacetted, and so can be expressed, tagged and approached in many different ways (Grieff, Croft and Turtle 1997). The greater flexibility of this Bayesian network model does, however carry with it the penalty that extensive experiment is required to determine the appropriate node weighting and input combination functions.

But what is significant is that though these three models differ at the formal level, their instantiations have much in common and their performance is very similar. This naturally reflects not only the fact that the same data are available for requests and documents but, more importantly, that very extensive testing and evaluation in the field as a whole have led to an understanding of the real nature of the document retrieval task and of the properties of the data that matter for this. Thus all three approaches use $TF^*CFW$-type weighting, combining term frequency (modulated by document length) with collection frequency. Similarly all place primary emphasis on simple terms and term coordination, not (nowadays) because more sophisticated document parsing is impracticable, but because this form of indexing is appropriate to the task.

These points lead naturally to the analysis of IR's relation to AI. To introduce this, by summing up my account of IR, the material points are:

- *statistically-based methods* work (subject to reasonable starting points in the user's request);
- as methods of defining and manipulating *information* they are weak methods;
- but they are suited to the task because the task itself (i.e. each individual instance) is

*under-specified*;

- thus it is sufficient to represent document (and request) content *indicatively* rather than substantively;
- further, it is desirable, since original language presentations (words, perspectives ...) vary, to support flexible *approximate* matching;
- thus it is helpful to exploit *redundancy* in representations;
- and, because retrieval deals with mass files, it is also useful to exploit *frequency data* for index keys as suggesting topic importance (again, also, in a redundant manner).

# 4 IR and AI

In surveying document retrieval, approached as just described, from an AI point of view, I shall first briefly 'locate' IR within AI; and then discuss the specific properties of IR with respect to AI's concerns with the nature of knowledge representation, reasoning, and learning. In the following section, 5, I will consider extensions to the basic document retrieval task. This leads to the concluding section summarising the lessons that AI might profitably learn from all these forms of information management. My intention here is not to be doctrinaire, or to engage in sterile disputes about what constitutes intelligence, but to suggest worthwhile lines of research.

## 4.1 IR as an AI task

As mentioned in the Introduction, the document retrieval task necessarily involves humans: it is they who seek information to be got by reading documents. This has the advantage of not demanding the impossible of systems, the disadvantage that humans may provide too little guidance for search.

Document retrieval may also be the precursor for further text analysis, as in the DARPA TIPSTER Programme (TIPSTER 1996), where information was extracted from relevant documents, or in the DARPA Automatic Summarising evaluation (Hand 1997), where retrieved documents are summarised to make it more convenient for users to reach decisions about likely relevance. But in these cases the situation is changed, in important ways.

In extraction the independent specification of what kind of information is to be extracted means that the prior selective retrieval is just a practical device for reducing text processing effort: there is no extra significance in a fact extracted from a putatively relevant document as opposed to a non-relevant one, though if facts are missed because their source document is not selected this is a problem. In this scenario the retrieval task has been redefined as a simple filtering one and, further, as a filtering task without the user participation in reading documents about a topic on which the classic case depends. This is not to suggest that TIPSTER-style tasks are not valuable: the important point is what query-document matching in the first step implies about the status and value of facts selected in the second. While there could be a real relationship, e.g. facts gain cachet from being extracted from topic-matching documents, in current implementations this is not the case. In other database construction cases, where tuples, say, are extracted from texts (Alshawi 1987) (Young and Hayes 1985), there is a tacit presumption that all input texts are legitimate sources. In either case, however, extraction differs from retrieval in the material respect that the original discourse is discarded rather than retained.

In the summarising case the relationship with retrieval is closer, though variably so depending on the summary's function, e.g. to be read as a content substitute for the full text, or just as a means of deciding whether to tackle the full text, etc. In the first case the original text is thrown away, as in extraction, in the second it is kept, as in retrieval. However irrespective of function, in summarising the user maintains their critical role in discovering, interpreting and applying document content, though this shifted now to the summary as a document and this is read against a complex background where the reductive relation between source and summary discourses is acknowledged. The 'themes' characterising retrieved document subclasses offered in the AltaVista Refine search option are similarly intended for the user. Other forms of further document processing illustrated by methods for file clustering and display (e.g. (Cutting et al. 1992)), creation of hypertext links etc (Salton and Allan 1993), are equally user-motivated, though they vary in the way they define new discourses and preserve old ones.

Classical document retrieval thus falls in the class of AI tasks that assist the human user but cannot, by definition, replace them. Relying on the human contribution does not, however, make the machine one trivial: giving the user the right source of information is a serious enterprise. Retrieval also, just because something to do with information as a semantic and linguistic commodity is involved, overlaps AI, even if it does not fall within AI, and some so-called intelligent agents using methods of the type described earlier (cf. (Maes 1994)) are very simple programs.

Being able to rely on the human user is one advantage in IR. The other is that the material it deals with has an important property: IR's natural primitives, words, are intrinsically semantic, unlike, say, pixels. Even allowing for the need to go below the word to stems, or above it to compound terms, words are clearly a good starting point for object characterisation and recognition. This applies even though words have multiple senses, since it is fair to assume that these are related simply by being senses of the same word, and conjoint matching supplies further (sufficient) sense resolution. However, much is still required, as indicated in the previous section, to put this raw semantic system input to effective work.

### 4.2 The AI perspective

Now looking at the more specific AI characterisation of IR, this is as follows.

#### Knowledge representation

IR's representation both of individual entities and the world of entities is clearly very weak: concept names are not normalised (beyond stemming), and descriptions are mere sets of independent terms without structure or economy, indeed they are deliberately uneconomical. Weighting adds some refinement to the basic scheme, but is again a relatively coarse means of individuating terms or whole descriptions. Concepts and topics, term and description meanings, are left implicit, and there is no attempt to make them explicit e.g. in propositional form. The relation between terms is only association based on co-presence, and the implicit relations between terms may be either terminological or assertional, with both serving, indifferently, in matching to promote precision or recall.

For example, in [ `Navaho Indian blanket manufacture` ], `Indian` is strictly redundant but reinforces matching on `Navaho` or may substitute for it, increasing the chances of selecting correct senses of `blanket` and `manufacture`; the relations between `Navaho`, `blanket` and `manufacture` are of different kinds but can equally help to retrieve documents that are about the topic by matching on either `Navaho` and `blanket` or `Navaho` and `manufacture`.

The additional set differentiation supplied by the use of relevance data does not change the nature of the representations themselves, only their bearings. Statistically-defined compound terms imply a refinement of the basic scheme, but only a very minor one within the overall framework.

As already mentioned, the lack of normalisation and weak structure is designed to allow, in a tolerant and hospitable manner, for the many slightly different ways in which the same *generic* topic may be expressed and approached via searching. The presumption is that there is no need to set prior constraints on how topics should be formulated because, as documents and requests already embody legitimate uses of natural language, what succeeeds in matching well will de facto satisfy the constraints there actually are. This is what the key notion of *request-based* document indexing means.

This style of representation is the opposite of the classical AI type, and has more in common with connectionist ones. But it is better geared than full-blown connectionist approaches to user interaction where representations may need to be inspected and thus should be comprehensible (though problems have been experienced here with weighting and its effects, especially with short requests, which e.g. users of Web engines have found puzzling, and which system designers have sought to offset (Clarke, Cormack and Tudhope 1997) (Rose and Stevens 1997)).

### Reasoning

In line with the style of representation, reasoning is also weak: it is an approximate, squashy enterprise looking at what is in common between descriptions and preferring one item over another because more is shared (whether as different words or, via weighting, occurrences of the same word). The system's lack of definite commitment is also reflected in the provision of ordered output, responding to the lack of certainty in the data, and as such acceptable to the user. Rule-based approaches, modern versions of Boolean searching, have been tried (Vickery et al. 1987), but are expensive without any material performance gain. The probabilistic network approach, that allows for more varied forms of search statement and matching condition, does not alter the basic style of reasoning since the motivation is the same underlying one, namely 'throw together what you can get'.

### Learning

It will be apparent that the feedback techniques are forms of learning, specifically supervised ones. These again are relatively weak methods, given the uncertainty associated with each key for positive (and also negative) cases, in the usual situation where few documents from a very large file have been examined. This applies even though the reason for examining the documents is good, i.e. that they have been retrieved with a high score, and though relevance refers to the relation that each document independently has to the request. Thus though relevance data reduces the intrinsic uncertainty of keys as topic indicators, any individual key cannot be taken too seriously. (For further applications of learning see below.)

## 5 IR extensions

Before drawing out the implications of IR for AI, I shall consider extensions or relatives of the document retrieval task, and how far the approach used for that applies elsewhere. Some of these tasks, like summarising, have been viewed as more obvious candidates for AI than document retrieval has; and there is not yet enough evidence that they can be done using the type of method applied to retrieval. But equally, there is no good experimental evidence that they cannot, and there is a salutary lesson in the experience of applying NLP to document retrieval where, despite repeated claims that it was required, it has not delivered superior performance. This suggests that in the absence of real evidence about the need for heavy-duty AI, automating these tasks should be approached with an open mind.

To begin the review, it should be noted that there is no problem in principle about applying the methods described to other types of natural language document than the papers, reports, news stories, etc hitherto assumed. Indeed Web engines like AltaVista already apply them to e.g. seminar lists, advertising material, email, etc. The same applies to heavily-structured documents, for instance forms, e.g. museum catalogue records. Again, it is possible to cut across given initial document boundaries, for instance to retrieve best-matching sub-documents or *passages*, whether statically or dynamically defined, in their own right and not just as selectors for full documents as previously mentioned (Kaszkiel and Zobel 1997).

The methods primarily developed for English have been applied to languages with different primitive units (like Chinese characters), e.g. within TREC, and to spoken documents as well as written ones (also within TREC). It is further possible in principle to use the methods for artificial language data, for instance large bodies of software. However while such applications may have been made, the effort required for performance evaluation on the necessary large scale means that there is a lack of reports about them. It is also possible in principle to use the same generic methods for other types of natural semantic keys, for instance citations (Salton 1963), and as in practice already, URLs and other proper names. The presumption is that as long as these keys, as well as being semantic, have Zipfian distributional characteristics, like natural language words, the methods described will apply to them too.

All of these variations on the basic document retrieval case are natural ones, and deserve more thorough exploration. In particular, for instance, where a two-stage retrieval process is involved, with an initial phase retrieving a subset of documents and a further stage for instance to retrieve passages from high-ranking documents, the status of the $CFW$ component of a term weighting formula may have to be reexamined since query terms may occur in many documents in a select subset.

But whether approaches of the type described can be used for other information management tasks is more interesting.

The most obvious, closely-related task is *filtering*, i.e. assigning documents from an incoming stream to users, rather than retrospectively searching an existing file. This has been explored within the TREC framework, but only, until very recently, in a distorted way. Thus the original TREC 'routing' task assumed standing user interest profiles, with queries developed by training on past document matches and relevance assessments; but the modified queries were then applied to a fixed collection to give ranked output in the usual adhoc searching style. These tests (cf (TREC 1992-1998)) showed that the probabilistic and other models described could give very high performance with relevance feedback, especially with well-formulated initial profiles. But there was no testing of the true filtering case where a yes/no decision has to be made for each incoming document, and where there is continuous modification of the profile in response to the succession of decisions, which may well thus reflect a changing rather than static user need.

The TREC-6 and TREC-7 Filtering Track experiments have been designed to capture the filtering task more effectively, while overcoming the non-trivial problems of ensuring that the comparative evaluation involved is sound (Hull 1998). Adapting the usual ranking retrieval methods to incorporate and apply a decision threshold is not completely straightforward, so the current research programme is of great interest.

The filtering task is just one example of *categorisation*, or document/text classification, which may be undertaken for many purposes, for instance to assign thesaurus descriptors (Biebricher et al. 1988), or to allocate documents in the filtering style but to broadly-defined user classes, as with the spoken message categorisation described by Rose (1991). Biebricher et al. use an explicitly probabilistic approach. Other work on categorisation, for instance Hayes work with newsfeeds (Hayes, Knecht and Cellio 1988) (Hayes 1992), has explored NLP within a rule-based framework, albeit using shallow NLP methods. Since statistical approaches have proved effective in some cases, evaluations comparing these with NLP-based ones are much needed.

Less closely related, but equally or perhaps more important tasks are those requiring a more explicit expression of document content, in the direction that leads to summaries as texts, i.e. discourses, in their own right. The essential element here is the shift, from a characterision of document content geared to selective retrieval from a mass of material, to the characterisation of a document in its own right.

Despite considerable effort, statistically-based methods for forming abstracts by extracting source-text sentences do not give very satisfactory results (Luhn 1958) (Paice 1990) (Brandow, Mitze and Rau 1995), though the fact that services using them, perhaps along with other techniques, are offered on the Web e.g. by British Telecom and Microsoft, suggests that they may have some practical utility, especially where the source document is also readily available for reference. Statistically-based methods when used to extract longer passages, e.g. paragraphs, and perhaps to link them as hypertext documents (Salton and Allan 1993), is potentially more promising, since the length of the extracts ensures locally-coherent text.

More importantly, summaries are required for many different purposes, for instance alerting, previewing, condensing (Sparck Jones 1998c). So while summarising defined as producing a coherent discourse capturing and conveying the significant information content of a source text is well beyong the reach of statistically-based methods, other weaker forms of summary that are accessible to statistical techniques may be quite adequate for some purposes. In the current DARPA summarising experiment (Hand 1997), the requirement is to produce summaries that are a good enough to determine whether their source full text is relevant to a user's information need. For this purpose, minimal summaries consisting of no more than lists of statistically-extracted phrases may be sufficent, e.g. (Dersy 1996). But while the general weak approach may be useful, there is still work to be done to develop specific models like the probabilistic one in a manner suited to summarising, whether for individual documents or perhaps for sets of documents (in a manner related to the AltaVista generation of term sets for themes characterising sets of retrieved documents), as well as to determine precisely the tasks and contexts for which such limited summaries are of value.

As noted, the solid experimental evidence for the degree to which the basic methods applied to document retrieval can be extended to other types of data or, especially, related tasks is so far relatively limited. But, given both that these techniques have proved much more effective than originally expected, and that there are other information management tasks closely related to document retrieval, the possibilities are well worth exploring. This view is reinforced by the experience of NLP where the importance of corpus data has been increasingly recognised and the data have been exploited both to 'colour' conventional grammars and to motivate the type of robust parsing used e.g. for information extraction (Appelt et al. 1993), as well as to enhance a broad-coverage lexicon (Briscoe and Carroll 1997) or develop application-specific ones. The broader implication is thus that it is worth asking what generic tasks, or what specific applications, which appear to require NLP can in fact be effectively done with strategies that are indicative of, or point to, information rather than actually capture and express it.

At the same time, recent technological developments present new challenges for the extablished methods of document retrieval. Thus information retrieval on the World Wide Web is a wholly new ball game, where system designers are faced with at least four major problems: vast scale, extreme file heterogeneity, lack of permanent access to end documents, and user inertia, all interacting with one another.

Current Web search engines typically apply at least some of the methods described earlier, but usually in a fairly adhoc way and in combination with a range of other devices. However it is far from clear that techniques shown to be effective with half a million documents, as in the main TREC tests, can be simply further scaled up: there may rather be a need to develop the methods substantially, just as there was in scaling up from test files of a few thousand abstracts to ones with hundreds of thousands of full texts. The current TREC Very Large Collections Track (Hawking and Thistlethwaite 1998) suggests that with non-minimal requests and a high-precision need, performance can scale up to some millions of full-text documents; but Web searching may not meet these conditions. There are particular difficulties, for the statistical approaches, in responding properly to cases where search terms have very high $CFW$ or high $TF$ in short texts (as may occur with announcements or advertisements).

But the major problem is short requests. As noted, the retrieval methods described rely on the user as the source of the information need, but observation shows that Web users average about two initial search terms. This is in general too small a base from which to develop a discriminating query, and also, as mentioned earlier, leads to problems with output acceptability when documents matching on just one term outrank those matching on two. URLs are a new type of search key that are very valuable when the user is already aware of potentially promising locations, but for normal topic searches even proper names may be poor selectors with such large files. With few or undiscriminating terms, the 'false drop' problem looms large, since even a very low percentage value can mean an absolutely large number of unwanted documents. Developing the retrieval models, and also designing interfaces to inveigle the user into providing more productive requests or to engage in feedback, are new research areas made more challenging by the difficulty and cost of evaluation tests.

## 6 Lessons from IR for AI?

Weak, but model-based, methods have demonstrated their value for one form of information management, namely document retrieval, and have begun to be explored, in promising directions, for others. The claim here is that they work because, in situations where information demand, and hence supply, is underspecified, the right strategy is to be broadly indicative, rather than aggressively analytic (as in decision trees); and, further, that the appropriate way of being indicative is to allow many small and individually ambiguous clues to combine and interact within whatever match of document to query is found. The paper's case is that there are important tasks that can be labelled 'information management' - finding, recovering, reminding, sorting, grouping, tagging, etc, arising in very different contexts, that are quite crude and can often be done in *sufficient-to-the-day* mode because they are fundamentally inexact. If we apply exact techniques in these situations we will come unstuck, getting no answers or unhelpfully restricted answers, and spending more time than we should on trying to be correctly exact. The analogy is with a table, which can stand on any three legs, and it does not matter which three, at any one time, which is good enough for that time; trying to make it stable on all four legs means it can only stand in one place and involves hours fiddling with wedges or saws to make it do this. [2]

Using multiple, weak, redundant clues is not in itself a novel idea. The point of the paper is the data and task context, of natural language, where it has productive use in describing and redescribing entities. It is not clear whether, and how far, such approaches apply where there are no obvious semantic primitives, e.g. in image processing; but wherever words are used to describe things, the techniques presented here may have a role in providing access to information. This could cover much more of AI than might be supposed, since it includes not only all the domains where natural language discourse is the primary material, but those where natural language is used in conjunction with other representations or objects, as with text accompanying images or graphics. More importantly, it could also include all of those cases where formalised, artificial languages are used for description and and communication, as in agent worlds where knowledge representation and reasoning have such forms as:

```
(obliged g1)
(forbidden g2 high)
...
```

or

```
def-situation 'accept-help-desk-work-s
  :acting 'help-desk-attendant
  :received '(request ...
```

or

```
bel agt1 (bel agt2 (price high))
  ...
```

It could further include much more abstract function specifications or programs. These language objects may be less 'squashy' than natural language ones, but there may still be scope for applying the techniques described. Thus in the second example above, there is some semantic redundancy in 'help' and 'attendant', and some frequency variation across the terms e.g. 'help' is more frequent than 'received', which could support selective retrieval or descriptive tagging of knowledge-base sections. The challenge is thus to identify applications that are right for the methods described, where we should not go for overkill in defining information but where we can make something useful happen without needing to understand precisely

[2] Thanks to Jacques Heyman's *The stone skeleton*, CUP, 1995, for this illustration.

how or why, and where, if we do not succeed in reaching the task goal first time, there is plenty of opportunity for recovery and adaptation.

**Acknowledgement**

I am very grateful to Stephen Robertson and Stephen Walker for collaboration in the earlier work on the development and testing of the probabilistic model to which this paper largely refers, and to Stephen Robertson for comments on the paper.

# References


Alshawi, H. (1987) *Memory and context for language interpretation.* Cambridge: Cambridge University Press.

Appelt, D.E. et al. (1993) FASTUS: a finite state program for information extraction from real-world text. *IJCAI-93. Proceedings of the Thirteenth International Joint Conference on Artificial Intelligence*, 1172-1178.

Belkin, N.J., Oddy, R.N. and Brooks, H.M. (1982) ASK for information retrieval. Part I. Background and theory. *Journal of Documentation*, 38, 61-71.

Biebricher, B., Fuhr, N., Lustig, G. Schwanter, M. and Knorz, G. (1988) The automatic indexing system AIR/PHYS - from research to application. *Proceedings of the 11th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*, New York: Association for Computing Machinery, 333-342.

Brandow, R., Mitzem K. and Raus, L.F. (1995) Automatic condensation of electronic publications by sentence selection. *Information Processing and Management*, 31, 675-685.

Briscoe, E.J. and Carroll, J. (1997) Automatic extraction of subcategorisation from corpora. *Proceedings of the Fifth Conference on Applied Natural Language Processing*, Somerset NJ: Association for Computational Linguistics, 356-363.

Clarke, C.L.A., Cormack, G.V. and Tudhope, E A. Relevance ranking for one to three term queries. *Proceedings of RIAO-97, Computer-Assisted Information Searching on Internet*, Centre de Hautes Etudes Internationales d'Informatique Documentaires, Paris, 388-400.

Cleverdon, C.W. (1967) The Cranfield tests on index language devices. *Aslib Proceedings*, 19, 173-192.

Cooper, W. (1995) Some inconsistencies and misidentified modelling assumptions in probabilistic information retrieval. *ACM Transactions on Information Systems*, 13, 100-111.

Cutting, D.R., Karger, D.R., Pedersen, J.O. and Tukey, J.W. (1992) Scatter/Gather: a cluster-based approach to browsing large document collections. *Proceedings of the 15th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*, New York: Association for Computing Machinery, 318-329.

Dersy, J. (1996) *Producing summary content indicators for retrieved texts*, MPhil Dissertation, Department of Engineering, University of Cambridge.

Fagan, J.L. (1989) The effectiveness of a non-syntactic approach to automatic phrase indexing for document retrieval. *Journal of the American Society for Information Science*, 40, 115-132.

Grieff, W.R., Croft, W.B. and Turtle, H. (1997) Computationally tractable probabilistic modelling of boolean operators. *Proceedings of the 20th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*, New York: Association for Computing Machinery, 119-128.

Hand, T.F. (1997) A proposal for task-based evaluation of text summarisation systems. In *Intelligent scaleable text summarisation* (Ed. I. Mani and M. Maybury), Proceedings of a Workshop, 1997, Somerset NJ: Association for Computational Linguistics.

Harman, D.K. (1997) Evaluation techniques and measures. *The Fourth Text REtrieval Conference (TREC-4)*, (Ed. D.K. Harman), Special Publication 500-236, National Institute of Standards and Technology, Gaithersburg, MD, 1996, A6-A14.

Harter, S.P. (1975) A probabilistic approach to automatic keyword indexing. Parts 1 and 2. *Journal of the American Society for Information Science*, 26, 197-206 and 280-289.

Hawking, D. and Thistlethwaite, P. (1998) Overview of TREC-6 very large collection track. *The Sixth Text REtrieval Conference (TREC-6)* (Ed. E.M. Voorhees and D.K. Harman), Special Publication 500-240, National Institute of Standards and Technology, Gaithersburg, MD, 93-106.

Hayes, P.J. (1992) Intelligent high-volume text processing using shallow, domain-specific techniques. In *Text-based intelligent systems* (Ed. P.S. Jacobs), Hillsdale NJ: Lawrence Erlbaum.

Hayes, P.J., Knecht, L.E. and Cellio, M.J. (1988) A news story categorisation system. *Proceedings of the Second Conference on Applied Natural Language Processing*, Somerset NJ: Association for Computational Linguistics, 9-17.

Hull, D.M. (1998) The TREC-6 filtering track: description and analysis. *The Sixth Text REtrieval Conference (TREC-6)* (Ed. E.M. Voorhees and D.K. Harman), Special Publication 500-240, National Institute of Standards and Technology, Gaithersburg, MD, 45-68.

Ide, E. (1968) New experiments in relevance feedback. In *Scientific Report ISR-14*, Cornell University. Reprinted as Chapter 16 in *The SMART retrieval system* (Ed. G. Salton), Englewood Cliffs, NJ: Prentice-Hall, 1971.

Kaszkiel, M. and Zobel, J. (1997) Passage retrieval revisited. *Proceedings of the 20th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*, New York: Association for Computing Machinery, 178-185.

Luhn, H.P. (1958) The automatic creation of literature abstracts. *IBM Journal of Research and Development*, 2, 159-165.

Maes, P. (1994) Agents that reduce work and information overload. *Communications of the ACM*, 37 (7), 30-40.

Maron, M.E. and Kuhns, J.L. (1960) On relevance, probabilistic indexing and information retrieval. *Journal of the ACM*, 7, 216-244.

Mitra, M., Buckley, C., Singhal, A. and Cardie, C. (1997) An analysis of statistical and syntactic phrases. *Proceedings of RIAO-97, Computer-Assisted Information Searching on Internet*, Centre de Hautes Etudes Internationales d'Informatique Documentaires, Paris.

Mooers, C.N. (1951) Zatocoding applied to mechanical organisation of knowledge. *American Documentation*, 2, 20-32.

Okapi (1997) Papers on Okapi, Special Issue of *Journal of Documentation*, 33, 3-87.

Paice, C.D. (1990) Constructing literature abstracts by computer. Techniques and prospects. *Information Processing and Management*, 26, 171-186.

van Rijsbergen, C.J. (1979) *Information retrieval.* 2nd Ed, London: Butterworths.

Robertson, S.E. (1977) The probability ranking principle in IR. *Journal of Documentation*, 33, 294-304.

Robertson, S.E. (1990) On term selection for query expansion. *Journal of Documentation*, 46, 359-364.

Robertson, S.E. and Walker, S. (1994) Some simple effective approximations to the 2 Poisson model for probabilistic weighted retrieval. *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, New York: Association for Computing Machinery, 232-241.

Rocchio, J.J. (1965) Relevance feedback in information retrieval. In *Scientific Report ISR-9*, Harvard University. Reprinted as Chapter 14 in *The SMART retrieval system* (Ed. G. Salton). Englewood Cliffs, NJ: Prentice-Hall, 1971.

Rose, D.E. and Stevens, C. (1997) V-Twin: a lightweight engine for interactive use. *The Fifth Text Retrieval Conference (TREC-5)* (Ed. E.M. Voorhees and D.K. Harman), Special Publication 500-238, National Institute of Standards and Technology, Gaithersburg, MD, 279-290.

Rose, R.C. (1991) Techniques for information retrieval from speech messages. *Lincoln Laboratory Journal*, 4(1), 45-60.

Salton, G. (1963) Associative document retrieval techniques using bibliographic information. *Journal of the ACM*, 10, 440-457.

Salton, G. (1975) *A theory of indexing.* Philadelphia, PA: Society for Industrial and Applied Mathematics.

Salton, G. (1995) Remarks at the Symposium on Thirty Years of Information Retrieval at Cornell, April 1995. Video recording, Department of Computer Science, Cornell University.

Salton, G. and Allan, J. (1993) Selective text utilisation and text traversal. *Proceedings of Hypertext-93*, Washington DC: Association for Computing Machinery, 131-144.

Salton, G. and Buckley, C. (1988) Term weighting approaches in automatic text retrieval. *Information Processing and Management*, 24, 513-523.

Salton, G. and Buckley, C. (1990) Improving retrieval performance by relevance feedback, *Journal of the American Society for Information Science*, 41, 288-297.

Sparck Jones, K. (1998) (1998a) What is the role of NLP in text retrieval? In *Natural language information retrieval* (Ed. T. Strzalkowski), Dordrecht: Kluwer, 1-24.

Sparck Jones, K. (1998) (1998b) Summary performance comparisons: TREC-2, TREC-3, TREC-4, TREC-5, TREC-6. In *The Sixth Text REtrieval Conference (TREC-6)* (Ed. E.M. Voorhees and D.K. Harman), Special Publication 500-240, National Institute of Standards and Technology, Gaithersburg, MD, B1-B8.

Sparck Jones, K. (1998) (1998c) Automatic summarising: factors and directions. In *Advances in automatic text summarisation* (Ed. I. Mani and M. Maybury), Cambridge MA: MIT Press, 1-12.

Sparck Jones, K. (1999) Further reflections on TREC. *Information Processing and Management* (in press).

Sparck Jones, K. Walker, S. and Robertson, S.E. (1998) *A probabilistic model of information retrieval: development and status.* Technical Report 446, Computer Laboratory, University of Cambridge.

Taube, M. (1955) Storage and retrieval of information by means of association of ideas. *American Documentation*, 6, 1-18.

TIPSTER (1996) *Advances in Text Processing. TIPSTER Program Phase II*, Materials from the Workshop, May 4-6, 1996. Sponsored by DARPA; distributed by Morgan Kaufmann, San Francisco.

TREC (1992-1998): D.K. Harman (Ed.) *The First Text REtrieval Conference (TREC-1)*, Special Publication 500-207, National Institute of Standards and Technology, Gaithersburg, MD, 1993; *... Second ... (TREC-2)*. SP 500-215, NIST, 1994; *... Third ...(TREC-3)*, SP 500-225, 1995; *... Fourth ... (TREC-4)*, SP 500-236, 1996; Voorhees, E.M. and Harman, D.K. (Eds.) *... Fifth ... (TREC-5)*, SP 500-238, 1997; *... Sixth ... (TREC-6)* (1997), SP 500-240, 1998.

Turtle, H.R. and Croft W.B. (1991) Evaluation of an inference network-based retrieval model, *ACM Transactions on Information Systems*, 7, 187-222.

Vickery, A. Brooks, H., Robinson, B. and Vickery, B. (1987) A reference and referral system using expert system techniques. *Journal of Documentation*, 43, 1-23.

Young, S.R. and Hayes, P.J. (1985) Automatic classification and summarisation of banking telexes. *Proceedings, Second Conference on Artificial Intelligence Applications*, New York: Institute of Electrical and Electronics Engineers, 402-408.