INFORMETRICS 87/88
L. Egghe and R. Rousseau (Editors)

DETERMINISTIC STRATEGIES FOR QUERY (RE)FORMULATION IN INFORMATION RETRIEVAL

Vijay V. RAGHAVAN

The Center for Advanced Computer Studies, University of SW Louisiana, Lafayette, LA 70504-4330

Abstract

Relevance feedback is a process whereby a user examines the documents selected by a retrieval system and provides feedback to the system as to their relevance. Such a feedback can then be used to formulate an optimal query with respect to the current information need of the user. This process of query (re)formulation can be based on probabilistic concepts, where Bayesian decision theory provides the framework for a decision rule, or ideas which instead employ *deterministic strategies*. The former class of techniques are limited by the fact that (strong) assumptions have to be made concerning the nature of the conditional probability density functions characterizing the data. In contrast, deterministic techniques, which do not require any explicit assumptions about the distribution of the various descriptor values, can be adopted. Such methods would have the advantages of being "non-parametric" and robust (useful in a wide variety of contexts). A class of deterministic techniques has been advocated by Salton and the SMART project group at Cornell. However, that approach does not obtain a new query that can be claimed to be optimal in a certain sense. In this work, a deterministic method that obtains an optimal query according to a prescribed criterion is advanced. Furthermore, it will be demonstrated that such methods are applicable not only when there are two classes of relevance (relevant and non-relevant) but also when the feedback distinguishes documents according to several degrees of relevance.

1. INTRODUCTION

Information retrieval (IR) systems are designed to provide references to documents that would contain the information desired by the user. A document in a collection (data base) is defined as relevant or non-relevant to a given user query depending on whether or not the document is judged by the user to have the desired information. Thus, in order to contend with the problem of distinguishing the relevant documents from the non-relevant ones, it is necessary to adopt methods that facilitate the ranking of documents in the order of their potential relevance.

Several kinds of models have been proposed in the literature for this purpose. The most popular among these, particularly in research investigations, is the Vector Space Model (VSM). A number of investigations, which demonstrate that the VSM is effective in ranking documents according to their estimated relevance, have been reported [11-13]. Other advantages of this model are that it is conceptually simple and computationally efficient. Furthermore, it has been possible to characterize and investigate, in a natural way, processes such as

relevance feedback [1,10,11,18,20], document-space modification [12,13], and discriminant value analysis [12,13], which play a crucial role in the overall search and retrieval scheme.

Among these, the process of relevance feedback is of particular importance. It is important due to the inherent difficulties associated with the representation of the content of a document and the inability of formal query languages to adequately represent user needs. In such an environment, in addition to the basic retrieval facilities, it is necessary to also provide mechanisms by which the system can adaptively learn the concept of relevance vis-a-vis a particular user need. The process is referred to as relevance feedback since it requires the user to give a judgement as relevant or non-relevant to each retrieved document and this relevance information is used to determine how further searching will proceed.

Earlier work on relevance feedback may be broken down into two categories : The first is the class of techniques that may be deemed deterministic. This approach was initially advocated by Salton and his co-workers and several variations of this have been implemented and tested within the SMART retrieval system [10,11]. The main idea here is that the query vector is modified by using vectors corresponding to relevant (non-relevant) documents to enhance (diminish) the importance of the query terms that are prevalent in relevant (non-relevant) documents. The other class of techniques, in contrast, are based on probabilistic principles. More precisely, given the relevance information with respect to a set of documents, a representation of the query that, in some sense, best distinguishes the relevant items from the non-relevant ones is sought. [4,8,9,15-17,19].

The probabilistic techniques have the limitation that rather strong assumptions concerning the nature of the conditional probability density functions that characterize the distribution of terms are needed. In addition, the quantity of relevance information available is often so smal that one faces many practical difficulties in estimating the necessary parameters. In contrast, the deterministic techniques proposed in the literat re do not require such explicit assumptions about term characteristics. However, they have mainly been justified on intuitive grounds and can not be claimed to be optimal according to some criterion.

In this work, we advance a deterministic approach that aims to exhibit at least some positive characteristics of both classes of techniques mentioned above. On the one hand, the assumption needed are expected to be weaker than those of the probabilistic approach and, on the other hand, it will be possible to prescribe optimization criterion according to which the new query, based on relevance information, can be deemed the best.

The ideas in this paper are best understood within the general framework of the VSM. Consequently, in the next section, the basic notions pertaining to the use of the vector space model in IR are reviewed. Then, in section 3, the proposed approach is developed and illustrated. In section 4, our proposal is compared with existing methods for effecting relevance feedback and some preliminary experimental observations are presented. The final section summarizes this work and identifies an important direction for future research.

2. THE GENERAL VSM FRAMEWORK

We discuss the details of the vector space model without being concerned, at first, about the restrictions of what is given in the actual data and how they are interpreted. The basic premise of the VSM is that the various IR objects are imagined as elements of a vector space. Let $t_1, t_2, \ldots, t_n$ be the terms used in order to obtain a description of the documents in a collection. Corresponding to each term, t_i, suppose there exists a vector $\vec{t}_i$ in the space. Without loss of generality, it is assumed that the t_i's are vectors of unit length. Since we do not wish to impose any restrictive assumptions, we consider the t_i's as being the generating set of the subspace of interest. That is, any IR object of interest is a

linear combination of the $\vec{t_i}$'s, but the $\vec{t_i}$'s do not necessarily constitute the set of basis vectors. In particular, let the documents $d_1, d_2, \ldots\ldots, d_p$ be given by vectors $\vec{d_\alpha} = (a_{\alpha 1}, a_{\alpha 2}, \ldots\ldots, a_{\alpha n})$, for $1 \leq \alpha \leq p$.
Then, we have

$$\vec{d_\alpha} = \sum_{i=1}^{n} a_{\alpha i} \vec{t_i} , \qquad (1)$$

where $a_{\alpha i}$ is known as the component of $\vec{d_\alpha}$ along $\vec{t_i}$, $1 \leq \alpha \leq p$ and $1 \leq i \leq n$. Mathematically, the $\vec{d_\alpha}$'s are still not uniquely defined unless we specify either an explicit representation of the term vectors or how the $\vec{t_i}$'s are "related". In this connection, we need to introduce te concepts of linear dependence and scalar product.

Definition 2.1 A set of vectors $\{\vec{v_1}, \ldots\ldots, \vec{v_k}\}$ is linearly dependent if there exist scalars $c_1, c_2, \ldots\ldots, c_k$, not all zero, such that

$$c_1\vec{v_1} + c_2\vec{v_2} + \ldots\ldots\ldots + c_k\vec{v_k} = \vec{0}$$

Definition 2.2 Given a vector space V, the scalar product, . , between any two vectors $\vec{u}, \vec{v} \in V$ is a mapping

$$\cdot : V \times V \rightarrow R$$

that satisfies certain axioms [6]. In particular

$$\vec{u}.\vec{v} = |\vec{u}| \; |\vec{v}| \cos\theta, \qquad (2)$$

where $|\vec{u}|$ and $|\vec{v}|$ are the lengths of the vectors $\vec{u}$ and $\vec{v}$ and θ is the angle between them, satisfies the necessary axioms and hence, is a scalar product.
In order to circumvent the need for explicitly defining the $\vec{t_i}$'s, the scalar product between $\vec{t_i}$ and $\vec{t_j}$, $1 \leq i,j \leq n$, may be assumed to be known. Let these values be represented as a symmetric, n X n matrix, G_t, where the $(i,j)^{th}$ element equals $\vec{t_i} \cdot \vec{t_j}$. By eqn. (2) and the fact that the term vectors are specified to be of unit length, these elements can be written as

$$\vec{t_i} \cdot \vec{t_j} = \cos\theta , \qquad (3)$$

where θ is the angle between $\vec{t_i}$ and $\vec{t_j}$.
If the set of vectors $\{\vec{t_1}, \ldots\ldots, \vec{t_n}\}$ is linearly dependent, then the representation of the d_α's, according to eqn. (1), is not unique since certain term vectors can be replaced by a linear combination of other term vectors. In contrast, the set of term vectors being linearly independent implies that the term vectors together form a basis for the vector space and that the representation in eqn. (1) is unique. Furthermore, the set $\{\vec{t_1}, \vec{t_2}, \ldots, \vec{t_n}\}$, is linearly independent if and only if G_t is non-singular [6]. If every pair of vectors $\{\vec{t_i}, \vec{t_j}\}$, for $i \neq j$, in the set $\{\vec{t_1}, \vec{t_2}, \ldots, \vec{t_n}\}$ is orthogonal, then G_t is an identity matrix. Since in this special case G_t is non-singular, the set of term vectors is linearly independent. Note, however, that the set of term vectors being linearly independent only implies that G_t is non-singular, not that $G_t = I$.
Given the concepts above, we can immediately define, for retrieval purposes, the computation of similarity between a document and a query. Let $q = (q_1, q_2, \ldots, q_n)$ be the representation of the query. More precisely, if q_i is the component of the query q along term vector t_i, then the query vector is

$$\vec{q} = \sum_{i=1}^{n} q_i \vec{t_i} \qquad (4)$$

From eqn. (1) and (4), we can define the scalar product between $\vec{q}$ and $\vec{d}_\alpha$ as

$$\vec{d}_\alpha \cdot \vec{q} = \sum_{i=1}^{n} \sum_{j=1}^{n} a_{\alpha i} q_j \vec{t}_i \cdot \vec{t}_j \qquad (5)$$

writing this in the form of matrix equation, we obtain

$$R_q = \vec{q} G_t A' \qquad (6)$$

where $a_{\alpha i}$ is the element in row i and column α of matrix A', the transpose of A, and $R_q = (\vec{d}_1 \cdot \vec{q}, \vec{d}_2 \cdot \vec{q}, \ldots, \vec{d}_p \cdot \vec{q})$. Thus, eqn. (6) represents the computation corresponding to the retrieval function regardless of whether the set of term vectors $\{\vec{t}_1, \vec{t}_2, \ldots, \vec{t}_n\}$ is a basis. Another point to note is that, in order to compute R_q, in addition to $\vec{q}$, A and G_t must also be known.

2.1. The Standard Vector Space Model

The general framework just described implies that the environment is such that documents are represented by terms. That is, if a vector can be associated with each term, then any document is given as a linear combination of the term vectors. A subtle difficulty here is that there is no specification as to how one, at the outset, arrives at term vectors. This difficulty is, however, not so serious for purposes of IR modeling, since our main interest is in carrying out the computation specified in eqn. (5). More specifically, it is sufficient to know the orientation of the term vectors relative to each other, as given by the $\vec{t}_i \cdot \vec{t}_j$'s. In essence, any explicit representation of a vector may be seen as incidental.

The next issue, as far as the realization of the VSM is concerned, is the mapping of values known in the physical problem to the model components identified in eqn. (5). In Raghavan and Wong [7], this aspect is referred to as the problem of interpretation. Before addressing this issue, let us look at what is typically assumed given.

It is common in IR to describe the content of each document by means of keywords or index terms. These are usually derived from the text or some surrogate (e.g. abstract) through a process known as indexing. In addition to the selection of terms to represent documents, it is common also to associate weights that reflect the importance of each term as an indicator of content of the documents to which it is assigned. This aspect is the focus of research activities dealing with term weighting or term weight assignment. The result of these processing stages is a document-by-term matrix, W, where the $(\alpha,i)^{th}$ element $w_{\alpha i}$ of the matrix corresponds to the importance of term i in document α, for $1 \leq i \leq n$ and $1 \leq \alpha \leq p$. Similarly, for a query q, the weights of different query terms can be determined depending on their ability to characterize the user need.

Given this, it is necessary to develop approaches that facilitate the ranking of documents in the order of their estimated usefulness relative to a specific user need. The most common implementation of the VSM, referred to here as the standard VSM, involves :

(a) the correspondence of the i^{th} index term to a vector $\vec{t}_i$, $1 \leq i \leq n$, and the specification that any element of the vector space can be expressed as a linear combination of the term vectors. Specifically, documents and queries can be represented as vectors in this space;

(b) the interpretation of $w_{\alpha i}$ as the component of the vector corresponding to document α along i^{th} term vector. Thus, any document is expressed as a linear combination of the term vectors;

(c) the assumption that term vectors are pairwise orthogonal. That is, the scalar product, $\vec{t}_i \cdot \vec{t}_j$, of any two (normalized) term vectors equals 1 if i = j and is 0 otherwise.

Applying these conditions to eqs. (1), (4) and (5), we obtain

$$\vec{d}_\alpha = \sum_{i=1}^{n} w_{\alpha i}\vec{t}_i, \tag{7}$$

and

$$\vec{q} = \sum_{i=1}^{n} q_i\vec{t}_i, \tag{8}$$

for $1 \leqslant i \leqslant n$ and $1 \leqslant \alpha \leqslant p$, and

$$\vec{d}_\alpha \cdot \vec{q} = \sum_{j=1}^{n}\sum_{i=1}^{n} w_{\alpha i} q_j \vec{t}_i \cdot \vec{t}_j, \tag{9}$$

where the desired similarity is dependent on $\vec{t}_i \cdot \vec{t}_j$, $1 \leqslant i,j \leqslant n$. By point (c) above, eqn. (3) reduces to

$$\vec{d}_\alpha \cdot \vec{q} = \sum_{i=1}^{n} w_{\alpha i} q_i \,. \tag{10}$$

Furthermore the restriction in point (c) implies that the set of term vectors $\{\vec{t}_1, \vec{t}_2, \ldots, \vec{t}_n\}$ forms a basis for the vector space of interest. Under the conditions assumed, it is easily seen that only $w_{\alpha i}$'s and q_i's for $1 \leqslant i \leqslant n$ and $1 \leqslant \alpha \leqslant p$, need be specified.

3. THE PROPOSED FORMULATION FOR RELEVANCE FEEDBACK

The general VSM framework depicted by eqs. (1)-(6) assumes that the query and the document components along the term vectors as well as the scalar product between terms are given. Then the computation in eqn. (6) is performed to obtain R_q, which is the scalar product of the query and the various documents. In other words, the L.H.S. of that matrix equation is the only unknown quantity. But this is not the only way in which to exploit eqn. (6). Depending on the specification of what is assumed given and certain circumstantial factors, other uses of eqn. (6) will also be valid. In the remainder of this section, a specific set of conditions under which eqn. (6) leads to the modeling of relevance feedback are identified and discussed.

For our formulation, the modeling process follows the same steps as that used to obtain eqn. (6). The point of departure is in the operational aspects of that equation. More specifically, instead of R_q being the unknown, it is assumed to be specified via feedback and the components of $\vec{q} = (q_1, q_2, \ldots, q_n)$ are considered to be the unknowns. Consider, again, eqn. (6) given as

$$R_q = \vec{q} G_t A' \,.$$

Rewriting the equation so that R_q and q appear as column vectors, we obtain

$$\begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ a_{p1} & a_{p2} & \cdots & a_{pn} \end{bmatrix} \begin{bmatrix} \vec{t}_1 \cdot \vec{t}_1 & \cdots & \vec{t}_1 \cdot \vec{t}_n \\ \cdot & & \cdot \\ \cdot & & \cdot \\ \vec{t}_n \cdot \vec{t}_1 & \cdots & \vec{t}_n \cdot \vec{t}_n \end{bmatrix} \begin{bmatrix} q_1 \\ q_2 \\ \cdot \\ \cdot \\ q_n \end{bmatrix} = \begin{bmatrix} \vec{d}_1 \cdot \vec{q} \\ \vec{d}_2 \cdot \vec{q} \\ \cdot \\ \cdot \\ \vec{d}_p \cdot \vec{q} \end{bmatrix} \tag{11}$$

If the product of the matrices A and G_t is denoted as P, then eqn. (11) becomes

$$Pq' = R_q', \qquad (12)$$

where $\vec{q}$ is assumed to be the unknown. As an aside, it is worth recalling from Raghavan and Wong [7] that the $(i,j)^{th}$ element of P, which is $\vec{d_i}.\vec{t_j}$, is the projection $\vec{d_i}$ on $\vec{t_j}$, since the $\vec{t_i}$'s are unit vectors.

In order to demonstrate that the framework above can be adopted for relevance feedback, several issues have to be addressed. First, let us consider the R.H.S. of eqn. (12). Since this information must be known in advance, it is important to specify the kind of feedback that the user is expected to provide. In the introduction to this paper, it is presumed that in the mind of the user (vis-a-vis a particular need) there exists a dichotomy that divides the collection of documents at hand into relevant and non-relevant sets. Thus, R_q may be a vector of 1's and 0's, indicating the relevance or otherwise of documents. while this is perhaps the easier thing to do from the user point of view, for technical reasons concerning how one might solve for q_i's, it may be preferable to have users specify finer distinctions (more degrees of relevance). In any case, at the conceptual level, it seems that the general case is more appropriate.

Bookstein and Cooper [3], in developing a general model of IR systems, suggest that the function of a retrieval system can be seen to be one of defining a weak ordering, with respect to a request, of the collection of documents.

Specifically, the retrieval mechanism is modeled as a function that defines a mapping from the set of documents to a set of retrieval status values (set of values defining the degrees to which a document may be predicted to be relevant to the request). Thus, the document collection is in essence partitioned into a number of subsets and those subsets are ordered relative to each other.

In a recent paper, Bollman and Wong [2] consider the retrieval problem from a measurement theoretic view point. The user's judgement is corresponded to a preference relation and the retrieval function modeling the retrieval mechanism of the system is treated as a measure that correctly reflects the preference relation.

In view of the discussion above, it is expected that the R.H.S. of eqn. (12) will, in fact, be the specification by the user of his/her preference relation. Then, the process, of solving for the components of $\vec{q}$, will correspond to the discovery (learning) of the retrieval function that adequately reflects the preference relation. Naturally, a special case of such a preference relation is one where the user has only two levels : relevant and non-relevant. Since the retrieval status values defined via the scalar product of document and query vectors may have a different number of levels (or degrees) compared to the number of grades in the preference relation, the solution of the computation problem obtained may not be uniquely defined. If both the number of distinct retrieval function values and the number of grades of the preference relation have an exact correspondence, the computational problem can be seen as one of solving a system of linear equations. Otherwise, on the basis of R_q and P, a number of equations and inequalities can be formulated and a solution to such a system will be required.

A second issue that arises in this framework is due to the fact that the system of equation/inequalities to be solved can have many feasible solutions. Consequently, it would be advisable to provide some criterion by which some feasible solutions are deemed better than others. When this aspect is taken into account, the problem of finding the q_i's will become one of the mathematical programming involving the optimization of a function subject to certain constraints [5].

It is worth nothing, however, that depending on the nature of R_q and the way in which the criterion function, mentioned above, is specified, several different classes of optimization problems will result. The algorithms to solve those problems may correspondingly vary considerably.

A final issue of interest pertains to the amount of relevance information available. Clearly, this amount is directly related to the number of documents examined by the user. Whereas eqs. (11)-(12) imply that R_q is known relative to the whole collection, such a situation may not be assumed to hold in practice. Consequently, the number of constraints derived will be determined by the preference ordering provided with respect to the set of retrieved documents.

Notationally, let us start with eqn. (11) and make the standard interpretation that W = A. Then we obtain

$$R_q' = WG_tq'.$$

Furthermore, if $\hat{W}$ is imagined to correspond to documents examined by the user prior to certain iteration (say, k) of the relevance feedback process and R_q' the corresponding feedback obtained, then we desire the "best" query, q^k, to be used in the next iteration. This situation may be expressed as

$$\hat{W}G_t(q^k)' = R_q'. \tag{13}$$

Clearly, the number of rows of $\hat{W}$ (corresponding to the number of documents examined) and the manner in which R_q' is specified will dictate the number of linear inequalities and/or equations. Depending on the specifics, there may or may not exist a solution for q^k. When a solution does not exist, it means that a linear or second order decision boundary that correctly distinguishes the relevant from the non-relevant items can not be found. Here, again, a suitable criterion function that characterizes the error associated with a decision boundary can be provided and this function may be optimized.
Next, a simple example is provided to illustrate the main aspects of the proposed framework :

Example 3.1

For the sake of simplicity, we assume that a special case of the context represented by eqn. (13) holds. That is, $G_t = I$. In addition, let $w_{\alpha i}$ be either 1 or 0 depending on whether or not the document d_α is represented by term t_i. For the example, we have contrived a situation where a linear decision boundary, that perfectly separates the relevant documents from the non relevant ones, exists. Consequently, a criterion function is not introduced. Suppose that the documents already examined by the user at some point are represented by the matrix $\hat{W}$, below :

$$\hat{W} = \begin{array}{c} \\ d_1 \\ d_2 \\ d_3 \\ d_4 \\ d_5 \\ d_6 \end{array} \begin{array}{c} \begin{array}{cccc} t_1 & t_2 & t_3 & t_4 \end{array} \\ \begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 1 \\ 1 & 0 & 1 & 1 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \end{bmatrix} \end{array}$$

Naturally, it is implied that the set of index terms used is $\{t_1,t_2,t_3,t_4\}$. Furthermore, let the feedback from the user be that d_1 and d_2 are relevant and the rest are not. Then, plugging into eqn. (18), we obtain

$$\begin{bmatrix} 1 & 1 & 0 & 1 \\ 1 & 1 & 0 & 1 \\ 1 & 0 & 1 & 1 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} q_1 \\ q_2 \\ q_3 \\ q_4 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix},$$

where $q^k = (q_1,q_2,q_3,q_4)$ is the unknown. Since it is only necessary to ensure that d_1 and d_2 are ranked higher than d_α, $3 \leqslant \alpha \leqslant 6$, it is sufficient to derive certain inequalities that achieve the desired result. It can be shown that only the following inequalities matter :

(i)
$$d_1 \cdot q^k > d_6 \cdot q^k \Rightarrow q_1 + q_2 > q_1 + q_3 \Rightarrow q_2 > q_3,$$

(ii)
$$d_1 \cdot q^k > d_4 \cdot q^k \Rightarrow q_1 + q_2 > q_2 \Rightarrow q_1 > 0,$$

(iii)
$$d_1 \cdot q^k > d_3 \cdot q^k \Rightarrow q_2 > q_3 + q_4 \Rightarrow q_2 - q_3 > q_4,$$

(iv)
$$d_2 \cdot q^k > d_4 \cdot q^k \Rightarrow q_1 + q_4 > 0 \Rightarrow q_4 > -q_1,$$

(v)
$$d_2 \cdot q^k > d_6 \cdot q^k \Rightarrow q_2 + q_4 > q_3 \Rightarrow q_4 > -(q_2 - q_3).$$

Inequalities (iii) - (v) can be combined to give

$$(q_2 - q_3) > q_4 > \max[-q_1, -(q_2 - q_3)] .$$

A particular solution for q^k satisfying the above is $q^k = (1,1,0,0)$, and the resulting rank values for documents $\{d_1,d_2, \ldots, d_6\}$ are, respectively, $\{2,2,1,1,0,1\}$.

4. DISCUSSION

4.1 Comparison to earlier work

A well known approach for relevance feedback is the one due to Salton and his co-workers [10,11]. The main idea here is that the original query is modified on the basis of relevance information. More specifically, the equation

$$Q' = Q + \alpha \sum_{r_i \in R} r_i - \beta \sum_{s_i \in I} s_i \qquad (14)$$

defines the computational process, where Q, r_i and s_i are vectors corresponding respectively to the original query, the i^{th} relevant and retrieved document, and the i^{th} non-relevant and retrieved document. The constants and are parameters to be determined empirically. Thus, the new query is defined as a linear combination of the vectors corresponding to the original query as well as the relevant and non-relevant documents retrieved. Intuitively, the weights of terms that mainly appear in relevant documents will be enhanced in the new query and, similarly, the weights of terms mainly occurring in non-relevant documents will be diminished. It is expected that the new query will be closer to relevant documents retrieved and farther away from the non-relevant and retrieved. As a result, the chances of retrieving additional relevant items should be enhanced.

More recently, another class of relevance feedback techniques have been advanced. These are based on probabilistic principles; in particular, Bayesian decision theory. In these cases, the strategy is one of constructing an optimal query rather than that of query modification. The optimality criterion is, for example, minimization of the average probability of error. A fundamental aspect of these approaches is that the nature of the class-conditional density functions are assumed known. The computational process is one of precisely characterizing these density function via parameter estimation. A decision criterion, for distinguishing between relevant and non-relevant documents, is then expressed in terms of these parameters. For example, assuming that each document $d = (d_1,d_2, \ldots, d_n)$ is a binary vector (i.e. $d_i = 1$ or 0 depending on whether it is indexed by the i^{th} term) and that the assignment of any term is independent of any other term being assigned, separately, within the class of relevant and within the class of non-relevant documents, it has been shown that an optimal ordering can be provided by computing

$$\sum_{i=1}^{n} w_i \cdot d_i ,$$

where $w_i = \log \frac{p_i (1-q_i)}{q_i (1-p_i)}$. The p_i's and q_i's are parameters characterizing the density functions :

$$p(d \mid R) = \prod_{i=1}^{n} p_i^{d_i} (1-p_i)^{1-d_i}$$

and

$$p(d \mid I) = \prod_{i=1}^{n} q_i^{d_i} (1-q_i)^{1-d_i} .$$

In this context, the optimal query may be presented by the vector
$Q = (w_1, w_2, \ldots, w_n)$.
The proposed formulation is different from the query modification technique, given by eqn. (14), in the sense that our aim is to find an optimal query according to a chosen criterion. That is, we will have a formal specification of exactly the sense in which the generated query is good. When compared to the probabilistic approaches, the proposed method has the advantage that assumptions concerning conditional density functions are not necessary. Furthermore, not having to make assumptions is important in the sense that there is no need to worry about which assumptions are realistic and which are not. Since parameter estimation is avoided, other problems such as those caused by sizes of samples being too small are not as severe.
Another difference, which is more or less a direct consequence of the lack of assumptions, is that the results will be applicable to a broader class of problem instances [2]. On the negative side, however, it appears that the gains are made at the expense of having to use more costly algorithms.

4.2 Progress on algorithm development

As indicated, an important direction for research concerns the precise formulation of the relevance feedback process as an optimization problem. Concurrently, it is necessary to ensure that efficient algorithms are available to solve the resulting problem. A particular question of interest in this context is whether the algorithm is incremental. More precisely, considering eqn. (13) given by

$$\hat{W} G_t(q^k)' = R_q' ,$$

we would prefer a strategy that obtains q^k incrementally from q^{k-1}, rather than from scratch.
Some preliminary investigation has been carried out in this direction. For this purpose, the algorithm proposed by Hooke and Jeeves [14] was adopted. As a convenient starting point, it was assumed that R_q can be specified precisely. Thus, one linear equation resulted from each query-document pair. Our experiments indicated that when fewer equations than the number of unknowns were involved, q^k could still be computed. Furthermore, as the number of equations increased, the solutions became closer to the correct one (the exact solution that would result if the number of unknowns equaled the number of equations). The fact that this algorithm is efficient and incremental suggests that our proposal has promise. Since R_q can not be precisely known, via feedback, an algorithm such as that by Hooke and Jeeves to be modified to deal with inequalities. This aspect is currently under investigation.

5. CONCLUSIONS

Relevance feedback is a process whereby a user provides a judgement as either relevant or non-relevant with respect to each retrieved document and this information is used by the system to determine how further searching should proceed. A novel approach, based on the vector space model, for accomplishing this task is proposed. It is shown that the proposed scheme has advantages over both probabilistic and deterministic techniques that are currently known. Future work is planned in the direction of designing efficient algorithms for the optimization problems that arise in this context.

REFERENCES

[1] Attar, R. and Fraenkel, A.S., "Local Feedback in Full-Text Retrieval Systems", Journal of the ACM, Vol. 24 (3), (1978) pp. 397-417.

[2] Bollmann, P. and Wong, S.KM., "Adaptive Linear Information Retrieval Models", Proceedings of the Tenth Annual International ACMSIGIR Conference on Research and development in Information Retrieval (1987), pp. 157-163.

[3] Bookstein, A. and Cooper, W., "A General Mathematical Model for Information Retrieval Systems", Library Quarterly, Vol. 46, (2), pp. 153-157.

[4] Croft, W.B., "Experiments with Representation in a Document Retrieval System", Information Technology : Research and Development, Vol. 2, (1), (1983), pp. 1-21.

[5] Duda, R.O. and Hart, P.E., Pattern Classification and Science Analysis, (Wiley, New York, 1973).

[6] Greub, W.H., Linear Algebra, (Academic, New York, 1963).

[7] Raghavan, V.V. and Wong, S.K.M., "A Criticial Analysis of Vector Space Model for Information Retrieval", J. of Amer. Soc. for Infor. Sci., Vol. 37, (5), (1986), pp. 279-287.

[8] Robertson, S.E. and Sparck Jones, K., "Relevance Weigthing of Search Terms", J. of Amer. Soc. Inform. Sci., Vol. 27, (1976), pp. 129-146.

[9] Robertson, S.E., Maron, M.E. and Cooper, W.S., "Probability of Relevance : A Unification of two Competing Models for Document Retrieval", Information Retrieval : Research and Development, Vol. 1, (1987), pp. 1-21.

[10] Rocchio, J.J. Jr., "Document Retrieval Systems - Optimization and Evaluation" Doctoral thesis. In : Information Storage and Retrieval, Scientific Report, ISR-10, (Harvard University, Cambridge, Mass., 1966).

[11] Salton, G., The SMART Retrieval System - Experiments in Automatic Document Processing, (Prentice-Hall, Englewood Cliffs, New Jersey, 1971).

[12] Salton, G., Dynamic Information and Library Processing, (Prentice-Hall, Englewood Cliff, New Jersey, 1975).

[13] Salton, G. and McGill, M.J., Introduction to Modern Information Retrieval, (McGraw-Hill, N.Y., 1983).

[14] Siddall, J.N. Optimal Engineering Design : Principles and Applications, (Marcel Dekker, Inc. New York and Basel, 1982).

[15] Sparck Jones, K., "Experiments in Relevance Weighting of Search Terms", Journal of Documentation, Vol. 35, (1979) p. 133-144.

[16] Sparck Jones, K., "Search Term Relevance Weighting Given Little Relevance Information", Journal of Documentation, Vol. 35, (1979), pp. 30-48.

[17] Van Rijsbergen, C.J., "A Theoretical Basis for the Use of Co-occurrence Data in Information Retrieval", Journal of Documentation, Vol. 33, (1977), pp. 106-119.

[18] Vernimb, V., "Automatic Query Adjustment in Document Retrieval", Information Processing and Management, Vol. 13, (6), (1977), pp. 339-353.
[19] Yu, C.T. and Salton, G., "Precison Weighting-An Effective Automatic Indexing Method", JACM, Vol. 23, (1976), pp. 76-88.
[20] Yu, C.T., Luk, W.S. and Cheung, T.Y., "A Statistical Model for Relevance Feedback in Information Retrieval", JACM, Vol. 23, (1976), pp. 273-286.