VECTOR SPACE MODEL OF INFORMATION RETRIEVAL - A REEVALUATION*

S.K.M. Wong
Department of Computer Science, University of Regina,
Regina, Sask. Canada, S4S 0A2

Vijay V. Raghavan+
Department of Computer Science, University of Regina,
Regina, Sask. Canada, S4S 0A2

Abstract. In this paper we, in essence, point out that the methods used in the current vector based systems are in conflict with the premises of the vector space model. The considerations, naturally, lead to how things might have been done differently. More importantly, it is felt that this investigation will lead to a clearer understanding of the issues and problems in using the vector space model in information retrieval.

1. INTRODUCTION

Information Storage and Retrieval (ISR) is a discipline involved with the organization, structuring, retrieval and display of bibliographic information. ISR systems are designed with the objective of providing, in response to a user query, references to documents which would contain the information desired by the user. A typical application for a computerized ISR system is in a library environment where the database consists of books, journals, etc.

It is common in information retrieval to represent each document by means of keywords or index terms. When a user submits a request, which is also specified in terms of the keywords, the request is compared with the document

---

* This research was supported in part by a grant from the Natural Sciences and Engineering Research Council of Canada.

+ This author is on leave from Univ. of Regina and is currently with Institut fur Informatik, TU Berlin, Franklinstr. 28/29, Sekr. FR 5-8, 1000 Berlin 10, FRG.

representations to determine which of the documents should be retrieved. In essence, then, the system must retrieve references that are of value, or relevant, to the user's request.

A document may or may not be relevant to a user query depending on many variables concerning the document (what it is about, how is it organized, is it clear, etc.) as well as on numerous user characteristics (the reason for search, previous knowledge, does the user know what he wants, etc.). Since relevance depends in a complex way on many factors, it is recognized that an ISR system cannot precisely select only and all relevant documents. It has, therefore, been suggested that a retrieval system should attempt to rank documents in the order of their potential relevance to a user query.

One approach which has been widely used over the years to provide such a ranking models documents and queries as vectors (Salton 1971; van Rijsbergen 1979; Salton & McGill 1983). The keywords used to describe the contents of documents or queries are assumed to correspond to the various elements of the vectors. Thus, if the indexing vocabulary consists of n distinct keywords, each document is an n-element vector in which the i^{th} element represents the importance of the i^{th} keyword to the document concerned. When a query is presented, the system formulates the query vector and matches against the documents based on a chosen method of determining similarity between vectors. For example, similarity between the query and a document may be defined as the scalar product of the corresponding vectors and the documents could be ranked in the decreasing order of this measure.

2. MOTIVATION

It is clear that the notions presented above are completely informal. Formal notions from linear algebra such as basis vectors, linear independence or dependence, and orthogonality are carefully avoided. Even the question of whether we have a vector space, that is, are the axioms to be obeyed by the elements of a vector space appropriate for

information retrieval, is not considered. In fact, it seems that in the early literature a conscious effort was made to not make any direct connection to vector spaces. Instead, a vector meant a tuple or what, in programming languages, is considered as a one-dimensional array of certain size.

Thus, the notion of vector, considered above merely refers to data structure; some sort of a physical or even simply a notational aspect. Similarly, the scalar product or some other similar similarity function is simply an operation defined on the data structure. The logical model was usually quite different. For example, information retrieval objects and processes have been modelled in set theoretic terms, as well as in statistical terms involving notions such as random variables and density function. The main point here is that the concept of a vector was not intended to be a logical or a formal tool. In fact, the earliest reference we have come across to the idea of vector spaces in information retrieval literature was in Salton et al.(1975). In this work a brief mention is made to the possibility of viewing index terms as corresponding to various dimensions of a space, and the documents as vectors in such a space. But subsequent developments in that paper do not really depend on the modelling of information retrieval objects as vectors in a vector space. Moreover, in several other papers, the term vector processing model is used, and we presume by conscious choice, instead of the term vector space model (Salton 1980; Salton et al. 1983). Given the "vector" model as outlined above, all things are fine and dandy and one felt quite content.

However, certain recent developments have aroused our curiosity to ask whether in the information retrieval context one should take the vector space model seriously. Salton and McGill (1983), van Rijsbergen (1979), and Koll (1979) make specific mention of vectors in a multidimensional vector space. Van Rijsbergen (p.41) infers that Salton considers document representatives as binary vectors embedded in an n-dimensional Euclidean space. Koll observes that in the SMART system environment a vector space, in which the various dimensions correspond to the different index terms in the

vocabulary and where the terms are mutually orthogonal, is assumed. Salton and McGill (p.129-130) discuss these ideas further and point out that the assumptions involved are only first order approximations to the true situation. First, it is felt that the vector view does not adequately capture the notion of scope. Secondly, treating each index term as a separate coordinate and assuming the terms as being orthogonal, is deemed contrary to the reality where term relationships exist and index terms are not assigned independently of each other. These statements warrant careful scrutiny.

The issue, at the outset, is not so much that it is not clear what precisely these statements mean. Rather it is which of the two notions of vectors we wish to adopt in information retrieval. On the one hand, we can accept the easier of the two answers: that all along the notion of vector was used only in the sense of a tuple or an array. This represents the easier answer since it appears to be consistent with most of the traditional and fairly well accepted practices in our field. It would mean, though, that any of the references that have been made to notions such as vector spaces, n-dimensional space, and orthogonality should be disregarded as casual flirtings and be not taken seriously. On the other hand, we may assert that all along the notion of vectors was intended as a logical construct and that information retrieval objects and processes can be understood in the context of vector spaces. If this had been the case, we should be able to demonstrate that the traditional practices and interpretations are either consistent with or reasonable approximations of what is correct under the vector space model. Unfortunately, we find that earlier work in information retrieval is, for the most part, not consistent with the vector space model.

In this paper we, in essence, point out the ways the traditional approaches are in conflict with the premises of the vector space model. The considerations, naturally, lead to how things might have been done differently. More importantly, it is felt that this investigation will lead to a clear understanding of the issues and problems in using the

vector space model in information retrieval.

In addition to the new insight one gains about the modelling of information retrieval objects, their relationships and processes, the current work is also significant in that it lays the groundwork for a model that is reminiscent of that used in the WEIRD system by Koll (1979). More specifically, both terms and documents are represented by being a combination (or "mean" location) of term vectors or concepts that they contain. Similarly, terms may be viewed as a combination of documents or concepts. It is also possible to investigate the problem of dimensionality and identify a (vector) space of fewer dimensions than the number of distinct index terms.

3. THE VECTOR SPACE MODEL

The basic premise of adopting the vector space model is that the various information retrieval objects are modelled as elements of a vector space. Specifically terms, documents, queries, concepts, and so on are all vectors in the vector space. The existence of a vector space implies that we have a system with the linear properties: the ability to add together any two elements of the system to obtain a new element of the system and the ability to multiply any element of the system by a real number. Furthermore, the vectors obey a number of basic algebraic rules or axioms (e.g. $\underline{x} + \underline{y} = \underline{y} + \underline{x}$, for any vectors $\underline{x}$, $\underline{y}$). Note that a letter with underscore denotes a vector.

Let us first consider the issue of representation of documents in terms of the index terms. Let $t_1, t_2, \ldots t_n$ be the terms used to represent documents. Corresponding to each term, t_i, there exists a vector $\underline{t}_i$ in the space. Without loss of generality, it is assumed that $\underline{t}_i$'s are vectors of unit length. Now, suppose that each document D_r, $1 \leq r \leq m$, is a vector expressed in terms of $\underline{t}_i$'s. Let the document vector $\underline{D}_r$ be

$$\underline{D}_r = (a_{1r}, a_{2r}, \ldots a_{nr}).$$

Since it is sufficient to restrict our scope of discussion to

the subspace spanned by the term vectors, the $\underline{t}_i$'s can be thought to be the generating set. Every vector in this subspace, and in particular all document vectors, are linear combinations of the term vectors. Thus, $\underline{D}_r$ can be, equivalently, expressed as

$$\underline{D}_r = \sum_{i=1}^{n} a_{ir}\underline{t}_i \ . \qquad (1)$$

The coefficients a_{ir}, for $1 \leq i \leq n$ and $1 \leq r \leq m$, are the components of $\underline{D}_r$ along the $\underline{t}_i$'s.

We next introduce one of the most important concepts in vector spaces, that of linear dependence. A set of vectors $\underline{Y}_1, \underline{Y}_2, \ldots \underline{Y}_k$ are linearly dependent if we find some scalars $a_1, a_2, \ldots a_k$, not all zero, such that

$$a_1\underline{Y}_1 + a_2\underline{Y}_2 + \ldots . a_k\underline{Y}_k = 0 \ .$$

Using several known theorems in linear algebra (Goult 1978), it can be seen that

(i) $\{\underline{t}_1, \underline{t}_2, \ldots \underline{t}_n\}$ being the generating set for our space implies that any set of linearly independent vectors in this space contains at most n vectors,

(ii) because a basis is a generating set consisting of linearly independent vectors, any basis of this space has at most n vectors and, hence, the dimension is at most n,

(iii) it is always possible to obtain a basis from a finite generating set by eliminating vectors dependent upon others,

(iv) given a basis, $\{\underline{t}_1, \underline{t}_2, \ldots \underline{t}_{n'}\}$, for $n' \leq n$, any vector $\underline{x}$ in the space has a unique expression of the form:

$$\underline{x} = \sum_{i=1}^{n'} c_i\underline{t}_i,$$

(v) if $\{\underline{t}_1, \underline{t}_2, \ldots \underline{t}_{n'}\}$ is a basis of our space, then any n' linearly independent vectors will form a basis, and

the dimension of the subspace is n'.

Thus, not only can documents be expressed as a linear combination of terms, but also terms as a linear combination of documents. The latter is true, of course, assuming there exists the necessary number of linearly independent documents. Notationally, if $\{\underline{D}_1, \underline{D}_2, \ldots \underline{D}_{n'}\}$ is a basis, each term, $\underline{t}_i$, has an expression of the form

$$\underline{t}_i = \sum_{r=1}^{n'} b_{ri}\, \underline{D}_r \ , \quad (i = 1,2,\ldots n) \ . \tag{2}$$

Clearly, we can also have documents expressed as a linear combination of a basis consisting of only documents. In fact, a basis could be made up of documents and terms mixed together because both of them are elements in the vector space.

Another important concept in this context is that of a scalar product. Given a vector space, V, by the scalar product $\underline{x}.\underline{y}$ of two vectors $\underline{x}, \underline{y} \in V$, we refer to the quantity $|\underline{x}||\underline{y}| \cos \theta$, where $|\underline{x}|$ and $|\underline{y}|$ are the lengths of the two vectors and θ is the angle between $\underline{x}$ and $\underline{y}$. It is easy to verify the usually required properties which specify how a scalar product interacts with the operations of addition, and multiplication by scalar, mentioned earlier (Goult 1978). A vector space equipped with a scalar product is called a Euclidean space.

The following definitions involving scalar products are well known:

(i) $|\underline{x}| = (\underline{x} \ . \ \underline{x})^{\frac{1}{2}}$,

(ii) any vector $\underline{x} \neq \underline{0}$ can be normalized; i.e. replaced by a proportional vector of unit length given by $\underline{x} \ / \ |\underline{x}|$,

(iii) $(\underline{x}/|\underline{x}|).\underline{y}$ is the projection of vector $\underline{y}$ onto the vector $\underline{x}$,

(iv) vectors $\underline{x}$ and $\underline{y}$ in a Euclidean space are orthogonal if $\underline{x}.\underline{y} = 0$,

(v) a basis such that the vectors are mutually orthogonal and each vector is normalized is called an orthonormal basis.

4. IMPORTANT CONCEPTS AND THEIR RELEVANCE TO EARLIER WORK IN INFORMATION RETRIEVAL

For reasons of clarity, in this section, it is assumed that the number of terms is equal to the dimension of the subspace of interest, and that the number of documents are exactly the same as the number of terms, i.e. $n'=n=m$. Recall, also, that the term vectors $\underline{t}_1, \underline{t}_2, \ldots \underline{t}_n$ are normalized. Furthermore, we assume the set of documents as well as the set of terms form a basis.

4.1 Computation of Correlations

From eqn. (1), we have

$$\underline{D}_r = \sum_{i=1}^{n} a_{ir}\underline{t}_i \quad , \quad (r = 1,2,\ldots n) \quad . \tag{3}$$

For any query q, the corresponding query vector has the expression

$$\underline{q} = \sum_{i=1}^{n} q_i \underline{t}_i \quad .$$

In the general case, the scalar product, which we suppose is the measure of correlation between two vectors, of $\underline{D}$ and $\underline{q}$ is

$$\underline{D}_r \cdot \underline{q} = \sum_{i,j=1}^{n} a_{ir} q_j \underline{t}_i \cdot \underline{t}_j \quad . \tag{4}$$

4.2 Projections vs. Components

Next we consider important relationships between components, projections, and the scalar products (vector correlations). By multiplying eqn. (3) by $\underline{t}_j$, (j=1,2,...n), on both sides, we obtain a system of linear equations:

$$\underline{t}_j \cdot \underline{D}_r = \sum_{i=1}^{n} a_{ir} \underline{t}_j \cdot \underline{t}_i \ , \qquad (j,r = 1,2,\ldots n). \qquad (5)$$

Since $\underline{t}_j$'s are unit vectors, the scalar product $\underline{t}_j \cdot \underline{D}_r$ is the projections of $\underline{D}_r$ onto $\underline{t}_i$. Eqn. (5) can be rewritten in a matrix form as follows:

$$P = G_t A \ , \qquad (6)$$

where

$$(P)_{jr} = \underline{t}_j \cdot \underline{D}_r \ ,$$

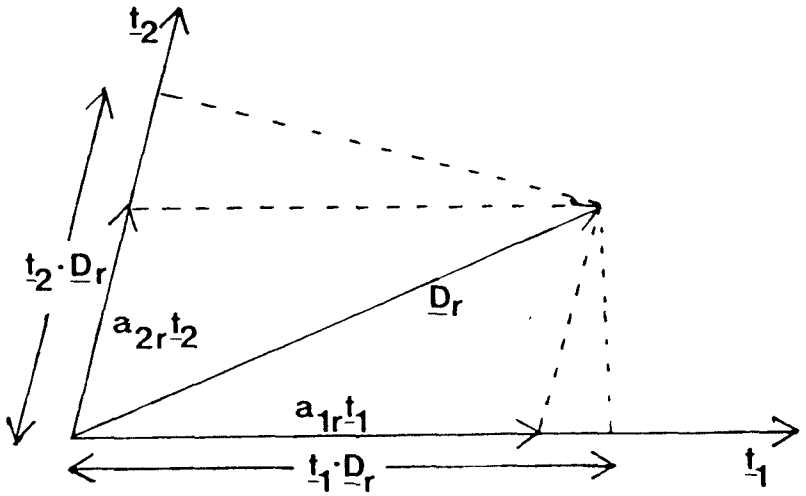
$$(G_t)_{ji} = \underline{t}_j \cdot \underline{t}_i \ , \quad \text{and}$$

$$(A)_{ir} = a_{ir} \ .$$

That is, G_t is the matrix of correlations between term vectors, and the r^{th} column of A represents the components of $\underline{D}_r$ along the vector $\underline{t}_i$'s.

Example 1. Consider a vector space with dimension n=2. In Fig.1 $\underline{t}_1$ and $\underline{t}_2$ represent the term basic vectors, and $\underline{D}_1$, $\underline{D}_2$ the document basic vectors.

Figure 1. Two Dimensional Vector Space with $\underline{t}_i$'s as Basis.



As in eqn. (3), each document vector $\underline{D}_r$ can be expressed as

$$\underline{D}_r = a_{1r}\underline{t}_1 + a_{2r}\underline{t}_2 \quad , \quad (r = 1,2).$$

The projection matrix P (defined in eqn. (6)) is given by

$$P = \begin{bmatrix} \underline{t}_1\cdot\underline{D}_1 & \underline{t}_1\cdot\underline{D}_1 \\ \underline{t}_2\cdot\underline{D}_1 & \underline{t}_2\cdot\underline{D}_2 \end{bmatrix} = \begin{bmatrix} \underline{t}_1\cdot(a_{11}\underline{t}_1+a_{21}\underline{t}_2) & \underline{t}_1\cdot(a_{12}\underline{t}_1+a_{22}\underline{t}_2) \\ \underline{t}_2\cdot(a_{11}\underline{t}_1+a_{21}\underline{t}_2) & \underline{t}_2\cdot(a_{12}\underline{t}_1+a_{22}\underline{t}_2) \end{bmatrix}$$

$$= \begin{bmatrix} \underline{t}_1\cdot\underline{t}_1 & \underline{t}_1\cdot\underline{t}_2 \\ \underline{t}_2\cdot\underline{t}_1 & \underline{t}_2\cdot\underline{t}_2 \end{bmatrix} \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} = G_t A \ .$$

If we multiply eqn. (3) by $\underline{D}_s$, (r = 1,2,...n), on both sides, we obtain

$$\underline{D}_s\cdot\underline{D}_r = \sum_{i=1}^{n} a_{ir}\underline{D}_s\cdot\underline{t}_i \quad , \quad (r,s = 1,2,...n),$$

which can be rewritten as

$$G_d = P'A \ , \tag{7}$$

where $(G_d)_{sr} = \underline{D}_s\cdot\underline{D}_r$ is the matrix of document correlations, and P′ is the transpose of P.

Similarly starting with eqn. (2), multiplying both sides by $\underline{D}_s$,(s = 1,2,...n), and $\underline{t}_j$, (j = 1,2,...n), respectively, we obtain the following matrix equations:

$$P' = G_d B \ , \tag{8}$$

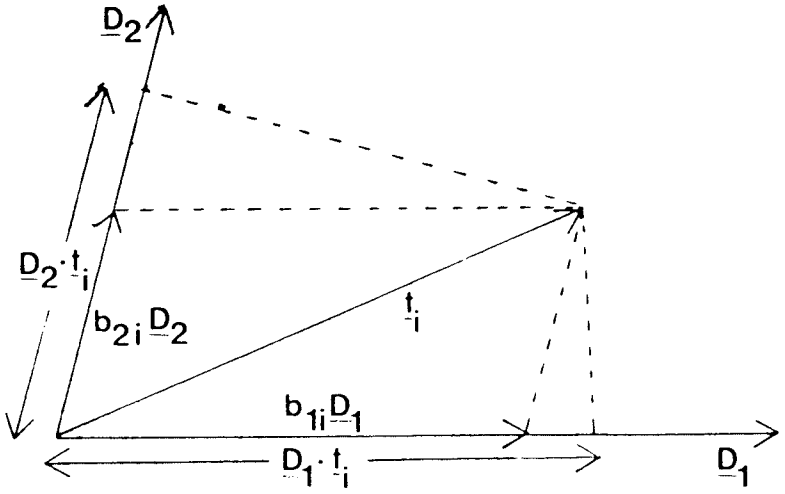
$$G_t = PB \ , \tag{9}$$

where $(B)_{ri} = b_{ri}$. The i^{th} column of B represents the

components of $\underline{t}_i$ along the directions of the various $\underline{D}_r$'s.

Example 2. Consider a two dimensional vector space as in Example 1. In this case the term vector $\underline{t}_i$ is expressed as a linear combination of the document basic vectors $\underline{D}_1$ and $\underline{D}_2$.

Figure 2. Two Dimensional Vector Space with $\underline{D}_r$'s as Basis.



4.3 Important Implications

Within the framework presented here, the assertions listed below will be established. This list is not intended to be exhaustive. In what follows, we let $\mathcal{D}$ denote the term-document matrix obtained from empirical data. That is, $\mathcal{D}$ is the matrix such that $(\mathcal{D})_{ir} = d_{ir}$, where d_{ir} is the occurrence frequency of term i in the document r.

(i) The model is usable in the general form as seen from the set of equations (6),(7),(8), and (9), if either (a) one of the correlation matrices (G_t or G_d) and one of A,B, or P are known, or (b) matrix P and one of A or B are known.

(ii) Recall from eqns. (6) and (9) that $G_t=PA^{-1}$ and $G_t=PB$. Therefore, in general, $B=A^{-1}$. Since $B=(G_d)^{-1}P'$, the correlation matrices G_t and G_d are related to each other as follows:

$$G_t = P(G_d)^{-1}P'$$

(iii) For the purpose of ranking documents against a query $\underline{q}$, it is clear from eqn. (4) that A and G_t must be known. We can, in fact, represent the scalar product $\underline{D}_r \cdot \underline{q}$, for r = 1,2,...n, as a vector $\underline{R}_q = (\underline{D}_1 \cdot \underline{q}, \underline{D}_2 \cdot \underline{q}, \ldots \underline{D}_n \cdot \underline{q})$, which can be written as

$$\underline{R}_q' = \begin{bmatrix} \underline{D}_1 \cdot \underline{q} \\ \cdot \\ \cdot \\ \cdot \\ \cdot \\ \cdot \\ \cdot \\ \underline{D}_n \cdot \underline{q} \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ \cdot & & & \cdot \\ \cdot & & & \cdot \\ \cdot & & & \cdot \\ \cdot & & & \cdot \\ \cdot & & & \cdot \\ \cdot & & & \cdot \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{bmatrix}' \begin{bmatrix} \underline{t}_1 \cdot \underline{t}_1 & \underline{t}_1 \cdot \underline{t}_2 & \cdots & \underline{t}_n \cdot \underline{t}_n \\ \cdot & & & \cdot \\ \cdot & & & \cdot \\ \cdot & & & \cdot \\ \cdot & & & \cdot \\ \cdot & & & \cdot \\ \cdot & & & \cdot \\ t_n \cdot \underline{t}_1 & \underline{t}_n \cdot \underline{t}_2 & \cdots & \underline{t}_n \cdot \underline{t}_n \end{bmatrix} \begin{bmatrix} q_1 \\ \cdot \\ \cdot \\ \cdot \\ \cdot \\ \cdot \\ \cdot \\ q_n \end{bmatrix}$$

$$= A' G_t \underline{q}' , \qquad (10)$$

where $\underline{R}_q'$ and $\underline{q}'$ (column vectors) denote the transpose of $\underline{R}_q$ and $\underline{q}$ respectively. Since G_t is a symmetric matrix and $P = G_t A$, eqn. (10) is equivalent to

$$\underline{R}_q = \underline{q}(G_t A) = \underline{q}P . \qquad (11)$$

If we assume that the term occurrence frequency d_{ir} represents the projection of the document vector $\underline{D}_r$ onto the term $\underline{t}_i$ (i.e. $\mathfrak{D} = P$), then eqn. (11) completely specifies the ranking of the documents with respect to the query $\underline{q}$ as follows:

$$(\underline{R}_q)_r = \sum_{j=1}^{n} q_j (D)_{jr} = \sum_{j=1}^{n} q_j d_{jr} , \quad (r = 1,2,\ldots n). \qquad (12)$$

It is important to note that there is no assumption made on term independence in the derivation of eqn. (12). Term correlations are implicitly included in the term occurrence frequencies d_{ir}'s by the assumption that $\mathfrak{D} = P$. This fact may explain why eqn. (12) works quite

well for document ranking in the traditional vector model. The advantage of interpreting $\mathcal{D}=P$ is that no explicit knowledge of term correlations (i.e. G_t) is required in eqn. (12) for computing correlations between the query and the documents. However, it is important to note that q_j's are the components (not projections) of $\underline{q}$ along the $\underline{t}_i$'s, and the matrix elements of A,B,G_t, or G_d are not known in this case.

(iv) The special case usually mentioned in the literature is obtained when $G_t=I$, i.e. the term vectors are assumed to be orthogonal and normalized, and $\mathcal{D}=A$. This means that the term occurrence frequency d_{ir} is interpreted as the component of the document vector $\underline{D}_r$ along $\underline{t}_i$. By eqn. (6), P=A which implies that the components of each $\underline{D}_r$ are identical to its projections onto $\underline{t}_i$'s. But $A \neq B$'! For this special case, eqn. (11) reduces to the well known form

$$\underline{R}_q = \underline{q}A = \underline{q}P = \underline{q}\mathcal{D} , \quad \text{or}$$

$$(\underline{R}_q)_r = \sum_{j=1}^{n} q_j a_{jr} = \sum_{j=1}^{n} q_j d_{jr} . \tag{13}$$

It is interesting to note that eqns. (12) and (13) give identical values for $(\underline{R}_q)_r$. However, contrary to case (iii), eqn. (13) is obtained by assuming explicitly that $\underline{t}_i$'s are orthogonal vectors. However, no such assumption is made in arriving at eqn. (12). Note that from eqn. (7) the matrix of document correlations is given by

$$G_d = P'A = \mathcal{D}'\mathcal{D} . \tag{14}$$

(v) A function commonly used to measure the similariy between term i and term j is

$$\sum_{r=1}^{n} d_{ir}d_{jr} \tag{15}$$

Clearly this function is akin to the well known term co-occurrence. Within the present framework, from eqn. (2) correlation between term vectors, $\underline{t}_i$ and $\underline{t}_j$, can be expressed as

$$\underline{t}_i \cdot \underline{t}_j = \sum_{r,s=1}^{n} b_{ri}b_{sj}\underline{D}_r \cdot \underline{D}_s \ . \tag{16}$$

If we assume that there exists no correlation between any pair of document vectors (i.e. G_d=I), then eqn. (16) becomes

$$\underline{t}_i \cdot \underline{t}_j = \sum_{r=1}^{n} b_{ri}b_{rj} \ . \tag{17}$$

From eqn. (8), the fact that G_d=I implies that P=B′. Hence, eqn. (17) can be rewritten as

$$\underline{t}_i \cdot \underline{t}_j = \sum_{r=1}^{n} (B')_{ir}(B)_{rj} = \sum_{r=1}^{n} (P)_{ir}(P')_{rj} \ , \tag{18}$$

or

$$G_t = PP' \tag{19}$$

If we further assume that $\mathcal{D}$ =P, term correlations can be computed directly from term co-occurrence frequencies as follows:

$$G_t = \mathcal{D}\mathcal{D}', \tag{20}$$

or

$$\underline{t}_i \cdot \underline{t}_j = \sum_{r=1}^{n} d_{ir} d_{jr} \cdot$$

Therefore, the interpretation of term co-occurrence frequency as term correlation has meaning in the vector space model only when document vectors are assumed to be orthnormal, and the term occurrence frequency d_{ir} is taken to be the projection of $\underline{D}_r$ along the term $\underline{t}_i$.

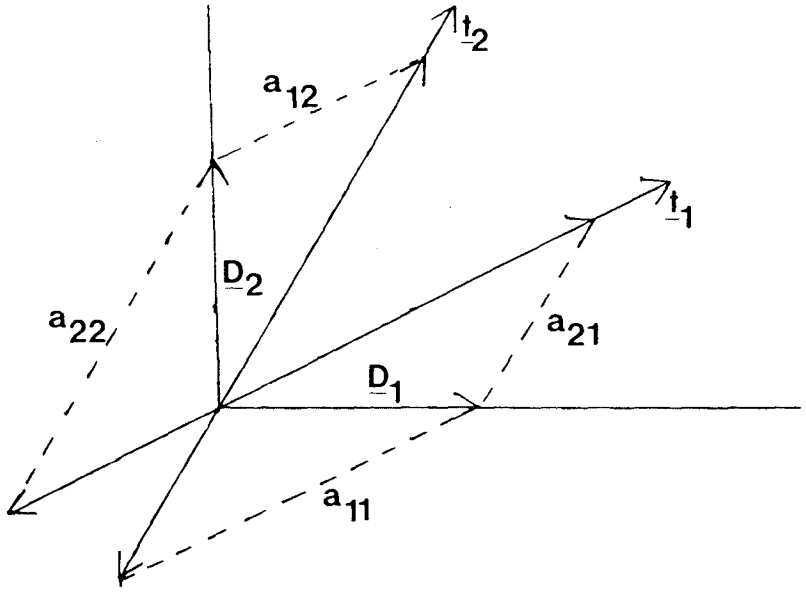
(vi) We may also assume that $G_t=G_d=I$ and $\mathfrak{D}=P$. Obviously from eqns. (6) and (8), we obtain $P=A=B'$. Since $G_t=I$, term co-occurence frequencies can no longer be interpreted as term correlations. The function defined by eqn. (15) is simply the scalar product, $\underline{t}_i \cdot \underline{t}_j$, evaluated with respect to the document co-ordinates frame of reference.

(vii) In view of the discussions presented in (iii), (iv), (v), and (vi), it is clear that given $\mathfrak{D}$ a decision has to be made as to what meaning to be attached to it. The interpretation that $\mathfrak{D}=P$ seems to be a reasonable one. If $\mathfrak{D}$ is assumed to be equal to A, the vector space model is not usable unless additional information is available or appropriate assumptions are made. However, if the matrix G_t is known, eqn. (4) shows precisely how term correlations should be incorporated into the retrieval strategy.

(viii) In current vector based models, all vector elements are conveniently assumed to be positive. Within the present framework, there is no reason to believe that all the matrix elements of A,B,P,G_t or G_d are necessarily positive numbers. In fact, both negative and positive vector elements are appropriate and necessary as can be seen from the following example.

Consider a two dimensional vector space shown below:

Figure 3. Negative Components in a Two Dimensional Vector Space.



The document vectors, $\underline{D}_1$ and $\underline{D}_2$, are expressed as a linear combination of the basic term vectors, $\underline{t}_1$ and $\underline{t}_2$:

$$\underline{D}_1 = a_{11}\underline{t}_1 + a_{21}\underline{t}_2 ,$$

$$\underline{D}_2 = a_{12}\underline{t}_1 + a_{22}\underline{t}_2 .$$

It can be seen from Fig. 3 that the component a_{21} of $\underline{D}_1$ along $\underline{t}_2$ and the component a_{12} of $\underline{D}_2$ along $\underline{t}_1$ are negative numbers. However, all the projections of the document vectors onto $\underline{t}_1$ and $\underline{t}_2$ are positive in this particular example. In this context, it is possible in a general case that term vectors may be negatively correlated. It may, therefore, be necessary in query processing to assume negative components for the query vector $\underline{q}$. Based on the argument presented here, the need to introduce negative term (document) correlations in the vector space model is apparent. It is also intuitively clear that two index terms may have "opposite" semantic meaning. We believe that the existing vector model fails to take into consideration

this important difference between index terms.

5. FURTHER ISSUES PERTAINING TO THE GENERAL VECTOR SPACE MODEL

At this point it may be concluded that the vector space model is inappropriate for information retrieval. Alternatively, one may investigate further to see what the issues and challenges are if one decided to model information retrieval by means of the vector space model in a more rigorous sense. In the latter direction, two important issues will be addressed:

(i) If $\mathfrak{D}$ can only represent A,B, or P, how might the additional information needed to fully specify the system be obtained. Based on the discussions in Section 4.3, the interpretation of the term occurrence frequency d_{ir} as the projection of $\underline{D}_r$ onto $\underline{t}_i$ (i.e. $\mathfrak{D}$ =P) seems to be a plausible one. In our view, the assumption that $\mathfrak{D}$ =P is, in fact, consistent with the well established practices in the vector based systems, provided that the document vectors, $\underline{D}_r$'s, are assumed to be orthonormal (i.e. G_d=I). We believe that to find a method for choosing an appropriate orthonormal basis for the document vectors is a crucial step in the development of a rigorous vector space model within the context of information retrieval. This task is currently under investigation by the authors of this paper.

(ii) Let $\{\underline{t}_1, \underline{t}_2, \ldots \underline{t}_n\}$ be a generating set of the term vector space. In order to have a unique expansion based on this set of $\underline{t}_i$'s for any vector in the vector space, a basis (i.e. a maximal subset of linearly independent term vectors) must be identified. Of course, this would be a trivial task if the term vectors were assumed to be orthogonal, because mutual orthogonality between vectors in a set implies linear independence. In contrast linear independence only implies that any redundancy in the usage of terms has been removed and the representation in terms of the resulting vectors is

compact (unique). Thus under non-orthogonality, correlation and dependence are rather distinct notions of term "relationship". The dimension of the space, clearly, ties in with these notions. In general, when the generating set of a vector space is not orthogonal, the task of identifying a basis may not be as trivial as it seems. First of all the issue of linear independence among an arbitrary set of vectors can be resolved only if correlation between any pair of the vectors is known. The approach we have suggested in (i) of this section may offer a solution to this problem with respect to term correlations. Secondly, even if term correlations are explicitly known, we still need a method for selecting a basis from the generating set. This problem is particularly troublesome when the number of term vectors is very large in practice. We have developed an algorithm for identifying a maximal subset of linearly independent term vectors provided that term correlations are known or approximated. The method will be reported in a forthcoming paper.

6. CONCLUSION

In our reevaluation of the vector space model, two main questions are raised:

(i) Whether the vector space model has been taken seriously in the information retrieval context?

(ii) Whether the vector space model should be taken seriously at all?

It is shown that in view of the well established practices, the answer to the first question is negative. However, based on our detailed analysis in Section 4, the answer to the second question is positive because it has been demonstrated that much insight can be gained by thinking about our problem within the framework of vector spaces. We believe that if the issues we have raised in Secion 5 can be satisfactorily resolved in the future, the vector space model looks very promising indeed and it will provide a useful and formal framework for the information retrieval systems.

Acknowledgement

We are grateful to Peter Bollmann, Ulrike Reiner and other members of the LIVE project group for helpful discussions on issues addressed in this paper.

References

Goult, R.J. (1978). Applied Linear Algebra. Chichester, England: John Wiley & Sons.

Koll, M. (1979). An approach to concept based information retrieval. ACM-SIGIR Forum, Vol. XIII (Spring), 32-50.

Salton, G. (1971). The SMART Retrieval System - Experiments in Automatic Document Processing. Englewood Cliffs, N.J.: Prentice-Hall.

Salton, G., C.S. Yang & A. Wong (1975). A vector space model for automic indexing. Comm. ACM, 18 (November), 613-620.

Salton, G. (1980). Automatic information retrieval. IEEE Computer, 13 (September), 41-56.

Salton, G. & M.H. McGill (1983). Introduction to Modern Information Retrieval. New York, N.Y.: McGraw-Hill.

Salton, G., E.A. Fox & H. Wu (1983). An automatic environment for boolean information retrieval. Information Processing '83. Proceedings of the IFIP 9th World Computer Congress, Paris, France, 755-762.

van Rijsbergen, C.J. (1979). Information Retrieval, 2nd Edition. Butterworth, London.