5,4 Dr. Raghavan 9

A Critical Analysis of Vector Space Model for Information Retrieval*

Vijay V. Raghavan and S. K. M. Wong
Computer Science Department, University of Regina, Regina, Saskatchewan, Canada S4S 0A2

Notations and definitions necessary to identify the concepts and relationships that are important in modelling information retrieval objects and processes in the context of vector spaces are presented. Earlier work on the use of vector model is evaluated in terms of the concepts introduced and certain problems and inconsistencies are identified. More importantly, this investigation should lead to a clear understanding of the issues and problems in using the vector space model in information retrieval.

1. Introduction

Information Retrieval (IR) is a discipline involved with the organization, storage, retrieval, and display of bibliographic information. IR 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.

Thus, in this environment there exists a collection of documents (e.g., books, journal articles, technical reports, etc.). There is also a group of users. The information need (at a particular time) of a user can be met by his reading one or more of the documents. The notion of relevance is central to any reasonable formulation of this retrieval problem. A document may or may not be relevant to a user query depending on many variables concerning the document (e.g., its scope, how it is written), as well as numerous user characteristics (e.g., why the search was initiated, user's previous knowledge). In any case, whatever the IR system does, if a document is judged by the user to be of interest, it is *relevant*; it is *nonrelevant* otherwise. Since many factors determine the judgement concerning relevance in a complex way, it is recognized that an IR system cannot precisely select only and all relevant documents. Rather, it is proposed that the system should adopt methods that facilitate the ranking of documents in the order of their estimated usefulness to a user query [1].

It is common in IR to represent 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 of *indexing*. In addition to the selection of terms to represent documents, it is common to also associate weights that reflect the importance of each term as an indicator of the content of the documents to which it is assigned. Thus, in designing search strategies, it is reasonable to consider a document-by-term matrix as the information one starts with, where the (i, r)th element of the matrix corresponds to the weight of term i in document r [2]. In what follows, we denote this matrix by $\mathfrak{D}$, having elements d_{ir}s.

Given the matrix $\mathfrak{D}$ and our desire to rank documents, there are several different ways to model the search problem. One approach which has been widely used over the years, models documents and queries as vectors [2,3]. Here, d_{ir} is considered to be the ith component of the vector representing the rth document. When a query is presented, the system formulates the query vector and matches it against the document vectors 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

In general, the use of vector space model requires the specification of several aspects. For example, an interpretation of the values in $\mathfrak{D}$, the dimension of the space, the set of basis vectors, and correlations between term vectors (if they are not orthogonal), are important in obtaining an appropriate representation of queries and doc-

*This research work is supported in part by an operating grant from the Natural Sciences and Engineering Research Council of Canada.

Received June 27, 1985; revised September 23, 1985; accepted November 14, 1985.

uments. In order to simplify the problems associated with the above requirements, other researchers adopting this model assume that:

(i) dimension of the vector space is n, where n is the number of distinct terms,
(ii) term vectors are pairwise orthogonal, and
(iii) d_{ir}s are the components of document r along the direction of term vector i [2].

Although the assumption that the terms are pairwise orthogonal is not realistic, it has been considered acceptable as a first approximation. In fact, many useful and interesting results have been obtained despite the simplifying assumptions [2–4].

While there may be good justification for starting an investigation with a simple model, one should not completely ignore the general case. In other words, if the general model is not clearly visualized in terms of the various concepts and interactions involved, there can be difficulties in not only recognizing the real implications of the special case, but also in subsequent attempts to deal with the general case. We believe that the way in which the vector space model has been introduced and used in IR has suffered difficulties of the sort mentioned above. For example, in order to relax the assumption that terms are pairwise orthogonal, term co-occurrence information has been used and some methods of computing term correlations suggest that the rows of $\mathfrak{D}$ can be viewed as vectors corresponding to the terms, i.e., $t_i = (d_{i1}, d_{i2}, \ldots, d_{in})$. However, we demonstrate in this paper that representing terms as rows of $\mathfrak{D}$ is not consistent with representing documents as the columns of $\mathfrak{D}$. In other words, operations and concepts needed for certain aspects were introduced without analysing their impact on other operations and concepts already in place. Such an inconsistency could not be resolved even under the restricted environment that is obtained when the assumptions mentioned earlier are made. In this sense, we believe that if the earlier work is not to be deemed inconsistent, then a way out is to say that the vector space model was not intended as a formal model of IR concepts and processes, but rather vectors were used as a notational convenience.

In order to establish these assertions, we introduce the notations and definitions necessary to identify the concepts and their relationships that are important for the use of vector space model in IR. Then we point out some of the ways in which the traditional practices 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 modeling 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 [5]. More specifically, both terms and documents are represented as 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 subspace 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 modeled 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 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., $x + y = y + x$, for any vectors x, y). Note that a letter by itself in italics denotes a vector. There are a few exceptions to this convention (e.g., m, n, and subscripts) and should be clear in context.

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, suppose there exists a vector t_i in the space. Without loss of generality, it is assumed that t_is are vectors of unit length. Now, suppose that each document D_r, $1 \leq r \leq m$, is a vector expressed in terms of t_is. Let the document vector D_r be $D_r = (a_{1r}, a_{2r}, \ldots a_{nr})$, where a_{ir}s are real numbers reflecting the importance of term i in D_r. Since it is sufficient to restrict our scope of discussion to the subspace spanned by the term vectors, the t_is 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, D_r can be, equivalently, expressed as:

$$D_r = \sum_{i=1}^{n} a_{ir} t_i. \qquad (1)$$

The coefficients a_{ir}, for $1 \leq i \leq n$ and $1 \leq r \leq m$, are the *components* of D_r along the t_is.

We next introduce one of the most important concepts in vector spaces, that of linear dependence. A set of vectors $y_1, y_2, \ldots y_k$ are *linearly dependent* if there exist some scalars $a_1, a_2, \ldots a_k$ (not all a_is are zero) such that:

$$a_1 y_1 + a_2 y_2 + \ldots + a_k y_k = 0.$$

Using several known theorems in linear algebra [6], it can be seen that:

(i) $\{t_1, t_2, \ldots t_n\}$ 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, $\{t_1, t_2, \ldots t_{n'}\}$, for $n' \leq n$, any vector x in the space has a unique expansion of the form:

$$x = \sum_{i=1}^{n'} c_i t_i,$$

(v) if $\{t_1, t_2, \ldots t_{n'}\}$ 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 $\{D_1, D_2, \ldots D_{n'}\}$ is a basis, each term, t_i, has an expression of the form:

$$t_i = \sum_{r=1}^{n'} b_{ri} D_r, \quad (i = 1, 2, \ldots n). \tag{2}$$

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* $x \cdot y$ of two vectors $x, y \in V$, we refer to the quantity $|x||y| \cos\Theta$, where $|x|$ and $|y|$ are the lengths of the two vectors and Θ is the angle between x and y. A vector space equipped with a scalar product is called a *Euclidean space.*

The following definitions involving scalar products are well known:

(i) $|x| = \sqrt{(x \cdot x)}$,

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

(iii) $(x/|x|) \cdot y$ is the *projection* of vector y onto the vector x,

(iv) vectors x and y in a Euclidean space are *orthogonal* if $x \cdot y = 0$,

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

4. Important Concepts and Relationships for Applying Vector Space Model in Information Retrieval

For reasons of clarity, in this and the next sections, 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 $t_1, t_2, \ldots t_n$ are normalized. Furthermore, we assume that the set of documents as well as the set of terms form a basis. Note, however, that the vectors in each set are not assumed to be pairwise orthogonal.

A. Computation of Similarity Measures between Documents and Query

From Eq. (1), we have

$$D_r = \sum_{i=1}^{n} a_{ir} t_i, \quad (r = 1, 2, \ldots n). \tag{3}$$

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

$$q = \sum_{i=1}^{n} q_i t_i.$$

In the general case, the scalar product, which we suppose is the measure of similarity between two vectors D_r and q, is:

$$D_r \cdot q = \sum_{i,j=1}^{n} a_{ir} q_j t_i \cdot t_j. \tag{4}$$

B. Projections versus Components

Next we consider certain important relationships between components, projections, and the scalar products. In the general case, where the basis vectors are not assumed to be pairwise orthogonal, components of documents along the term vectors are related to the corresponding projections via the term-term similarities. By multiplying Eq. (3) by t_j, $(j = 1, 2, \ldots n)$, on both sides, we obtain a system of linear equations:

$$t_j \cdot D_r = \sum_{i=1}^{n} a_{ir} t_j \cdot t_i, \quad (j, r = 1, 2, \ldots n). \tag{5}$$

Since t_js are unit vectors, the scalar product $t_j \cdot D_r$ is the projections of D_r onto t_j. Eq. (5) can be rewritten in a matrix form as follows:

$$\mathrm{P} = \mathrm{G}_t \mathrm{A}, \tag{6}$$

where

$$(\mathrm{P})_{jr} = t_j \cdot D_r,$$

$$(\mathrm{G}_t)_{ji} = t_j \cdot t_i, \text{ and}$$

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

That is, G_t is the matrix of correlations* between term vectors, and the rth column of A represents the components of D_r along the vector t_is.

Example 1. Consider a vector space with dimension $n = 2$. In Figure 1, let t_1 and t_2 represent the term basis vectors, and D_1, D_2 the document basis vectors. As in Eq. (3), each document vector D_r can be expressed as:

$$D_r = a_{1r}t_1 + a_{2r}t_2, \quad (r = 1, 2).$$

The projection matrix P [defined in Eq. (6)] is given by

$$P = \begin{bmatrix} t_1 \cdot D_1 & t_1 \cdot D_2 \\ t_2 \cdot D_1 & t_2 \cdot D_2 \end{bmatrix}$$

$$= \begin{bmatrix} t_1 \cdot (a_{11}t_1 + a_{21}t_2) & t_1 \cdot (a_{12}t_1 + a_{22}t_2) \\ t_2 \cdot (a_{11}t_1 + a_{21}t_2) & t_2 \cdot (a_{12}t_1 + a_{22}t_2) \end{bmatrix}$$

$$= \begin{bmatrix} t_1 \cdot t_1 & t_1 \cdot t_2 \\ t_2 \cdot t_1 & t_2 \cdot t_2 \end{bmatrix} \cdot \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} = G_t A. \quad \square$$

Similar to Eq. (6), a number of other relationships can be identified between document-document similarities, the projections of documents along terms, and certain other quantities inherent to the model. In what follows, we highlight the more important connections among those.

If we multiply Eq. (3) by D_s, $(r = 1, 2, \ldots n)$, on both sides, we obtain

$$D_s \cdot D_r = \sum_{i=1}^{n} a_{ir} D_s \cdot t_i, \quad (r, s = 1, 2, \ldots n),$$

which can be rewritten as

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

where $(G_d)_{sr} = D_s \cdot D_r$ is the matrix of (unnormalized) document correlations, and P′ is the transpose of P.

Similarly starting with Eq. (2), multiplying both sides by D_s, $(s = 1, 2, \ldots n)$, and t_j, $(j = 1, 2, \ldots n)$, respectively, we obtain the following matrix equations:

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

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

where $(B)_{ri} = b_{ri}$. The ith column of B represents the components of t_i along the directions of the various D_rs.

Example 2. Consider a two-dimensional vector space as in Example 1. In this case, the term vector t_i is expressed as a linear combination of the document basic vectors D_1 and D_2. Again, the transpose of the projection matrix, P, can be rewritten to illustrate the meaning of Eq. (8)

$$P' = \begin{bmatrix} t_1 \cdot D_1 & t_2 \cdot D_1 \\ t_1 \cdot D_2 & t_2 \cdot D_2 \end{bmatrix}$$

$$= \begin{bmatrix} (b_{11}D_1 + b_{21}D_2) \cdot D_1 & (b_{12}D_1 + b_{22}D_2) \cdot D_1 \\ (b_{11}D_1 + b_{21}D_2) \cdot D_2 & (b_{12}D_1 + b_{22}D_2) \cdot D_2 \end{bmatrix}$$

$$= \begin{bmatrix} D_1 \cdot D_1 & D_2 \cdot D_1 \\ D_1 \cdot D_2 & D_2 \cdot D_2 \end{bmatrix} \cdot \begin{bmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{bmatrix}$$

$$= G_d B. \quad \square$$

5. Vector Space Model in Information Retrieval—Choices and Implications

Given the concepts and notations in the earlier sections, we are now prepared to discuss a spectrum of op-

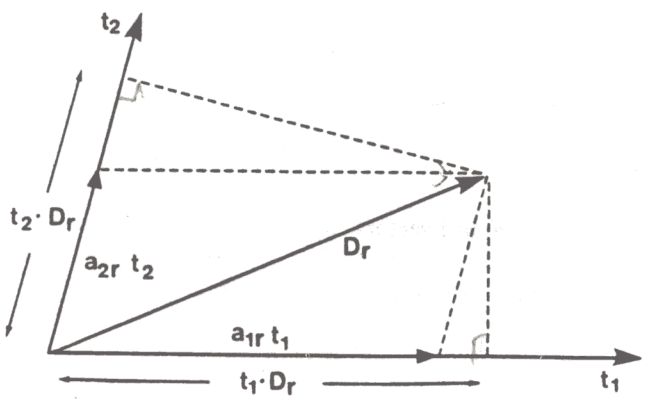

FIG. 1. Two-dimensional vector space with t_is as basis.

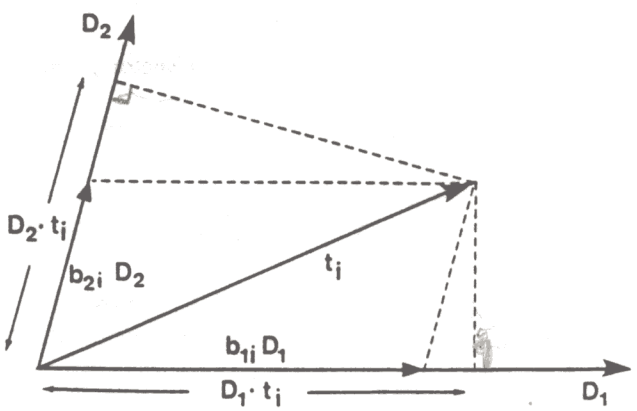

FIG. 2. Two-dimensional vector space with D_rs as basis.

*The word correlation is not intended to mean statistical correlations. The elements are term-term similarity measures. The values would range between −1 to +1, with zero value corresponding to the case when term vectors are orthogonal to each other.

tions available in terms of how the model can be applied. In the presentation of the vector space model (Section 3), our interest in modeling IR objects and processes was immediately recognized. However, we did not concern ourselves with what information would be available, and to what element of the model that information would be mapped. In other words, we basically looked at the kinds of concepts and relationships that one would have to work with. In this sense, the earlier discussions are abstract.

Some of the model elements identified are the component matrices A, B, the projection matrix P, and the term-term and document-document correlation matrices G_t and G_d. Since these are related, not all of these matrices need be known. Thus, one question that arises is, what should we know as a minimum? By a minimum, we mean that we must at least be able to rank documents in the order of their similarities to a given request. That is, the computation specified in Eq. (4) is of primary interest. Another issue that is equally important, once we specify what data we have about our documents and queries, is that of deciding which of that data should be mapped to what element of the model. This latter issue can be thought of as one of *interpretation.*

As in Section 4, we assume that $D_1, D_2, \ldots D_n$ as well as $t_1, t_2, \ldots t_n$ are (separately) linearly independent sets of vectors that span the subspace of interest. This is needed merely for ease of exposition since, in linear algebra, there is a well known theorem which says that a maximal linearly independent set can always be selected from a linearly dependent set [6]. We will, however, consider the case of a linearly dependent set of vectors in the next section.

A. Only Term Frequency Data is Known

In the introduction it was mentioned that it is common in IR environment to know just the occurrence frequencies of each index term in the documents of a collection. Let $\mathfrak{D}$ refer to the matrix of term frequencies, where the (i, r)th element, d_{ir}, is the frequency of occurrence of term i in document r.

Assuming that this and the query representation are all we know, the following option comes to mind.

(1) The Standard Vector Model. $\mathfrak{D}$ is interpreted to correspond to the component matrix A (i.e., d_{ir} is a_{ir}, the component of D_r along t_i). A problem that is well recognized is that knowing matrix A alone is not enough. Specifically, referring to Eq. (4), we cannot determine document-query correlation since $t_i \cdot t_j$s are not known. Therefore, in the standard vector model the assumption that $t_i \cdot t_j = 1$ if $i = j$, and 0 otherwise, is made (in other words, this assumption means $G_t = I$).

From Eq. (6), it follows that if $G_t = I$ then $P = A$. Thus, the above specifications imply the interpretation that $\mathfrak{D} = A = P$. That is, d_{ir}s are both projections and components of the documents along the terms vectors. Under these conditions,

$$D_r \cdot q = \sum_{i,j=1}^{n} a_{ir} q_j t_i \cdot t_j$$

$$= \sum_{j=1}^{n} a_{jr} q_j = \sum_{j=1}^{n} d_{jr} q_j, \qquad (10)$$

where $q = (q_1, q_2, \ldots, q_n)$ is the query vector and q_js are the components of q along the term vectors. The above form (dot product) of similarity function is, of course, well known in IR literature. This approach results in a model that is simple to apply and yet very useful. The problem here is that we no longer remain within the framework of this special case if we want to deal with term-term correlations.* However, document-document correlations can be obtained by

$$G_d = P'A = \mathfrak{D}'\mathfrak{D}.$$

(2) An Alternative Interpretation. In the standard vector model, G_t is assumed to be an identity matrix which then leads us to the interpretation that $\mathfrak{D} = P$. But, we can obtain the result equivalent (numerically) to Eq. (10) by a different interpretation of $\mathfrak{D}$. We achieve this by noticing that P is, in fact, a function of G_t and A.

To see this, let us represent the document-query similarity $D_r \cdot q$, for $r = 1, 2, \ldots n$, as a vector $R_q = (D_1 \cdot q, D_2 \cdot q, \ldots D_n \cdot q)$, which can be written as

$$R_q' = \begin{bmatrix} D_1 \cdot q \\ \vdots \\ D_n \cdot q \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} \ldots a_{1n} \\ \vdots \\ a_{n1} & a_{n2} \ldots a_{nn} \end{bmatrix} \begin{bmatrix} a_{11} \ldots a_{n1} \\ a_{12} \ldots a_{n2} \\ \vdots \\ a_{1n} \ldots a_{nn} \end{bmatrix}$$

$$\begin{bmatrix} t_1 \cdot t_1 & t_1 \cdot t_2 \ldots t_1 \cdot t_n \\ \vdots \\ t_n \cdot t_1 & t_n \cdot t_2 \ldots t_n \cdot t_n \end{bmatrix} \begin{bmatrix} q_1 \\ \vdots \\ q_n \end{bmatrix}$$

$$R_q = A'G_tq', \qquad (11)$$

where R_q' and q' denote the transpose of matrices R_q and q, respectively. Since G_t is a symmetric matrix and $P = G_tA$, Eq. (11) is equivalent to

$$R_q = q(G_tA) = qP. \qquad (12)$$

If we assume that the term occurrence frequency d_{ir} represents the projection of the document vector D_r onto the term t_i (i.e., $\mathfrak{D} = P$), then Eq. (12) completely specifies

*If the basis set $\{t_1, t_2, \ldots t_n\}$ is a proper subset of the set of all term vectors, we may find the terms not in the basis to be correlated with each other as well as with the terms in the basis set.

the ranking of the documents with respect to the query q as follows:

$$(R_q)_r = \sum_{j=1}^{n} q_j(\mathfrak{D})jr = \sum_{j=1}^{n} q_j d_{jr}, \quad (r = 1, 2, \ldots n). \tag{13}$$

Clearly, Eq. (13) is equivalent to the ranking given by Eq. (10). The important point is that this is also an interpretation which can be used to explain why the dot product of vectors q and D_r provide a measure of similarity. The columns of $\mathfrak{D}$ are, however, not document representations. Furthermore, term correlations are implicitly taken into account and they need not be known or assumed explicitly. The q_js are still interpreted as the components.

A disadvantage of this interpretation is that even if term-term correlations can be somehow obtained, the computation for document ranking will still be given by Eq. (13). Since empirical investigations show that we can do better than using the dot product, we believe that this interpretation although interesting will not lead to the development of better retrieval strategies.

(3) The Dual of the Standard Vector Model. In Section 5A(1), we discussed the most popular use of the vector space model in IR. We can instead interchange the role of documents and terms to obtain a new interpretation. Now suppose that $\mathfrak{D}$ is interpreted as B′, the components of terms along document vectors. Then,

$$t_i \cdot t_j = \sum_{r,s=1}^{n} b_{ri} b_{sj} D_r \cdot D_s. \tag{14}$$

Furthermore, let G_d be assumed to be an identity matrix. Under these conditions, Eq. (14) becomes

$$t_i \cdot t_j = \sum_{r=1}^{n} b_{ri} b_{rj}$$

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

For this special case, $P = B' = \mathfrak{D}$ implying that

$$G_t = PP'. \tag{16}$$

Equation (15) is interesting in that it is akin to well known term-term co-occurrence computations. We see that computation of term correlations in this way is valid only if $\mathfrak{D}'$ is interpreted to mean B.

From Eqs. (6) and (9) it follows that $A = B^{-1}$. Thus, after inverting B to obtain A, document-query correlations can be computed using A, and G_t as defined in Eqs. (15) or (16). It is important to note that in the standard vector model $\mathfrak{D} = A$, but here $\mathfrak{D} = B'$. We have to make one choice or the other but not both. In practice one finds the use of both Eqs. (10) and (15) at the same time, implying that $B^{-1} = A = B'$. The developments here show otherwise. In fact, to have $A = B'$ we must assume both $G_d = I$ and $G_t = I$. This special case is rather uninteresting, since we can neither talk about term-term correlations nor document-document correlations.

B. Use of Additional Information

From the discussions of Section 5A we realize that, given $\mathfrak{D}$, a decision has to be made as to what meaning we want to attach to it. Specifically, it seems that we have the option of interpreting $\mathfrak{D}$ as A, B′, or P.

It is also clear that, if we do not want to make any assumptions, we require additional information. Formally, to completely specify the objects and relationships, we must know either

(a) one of the correlation matrices (G_t or G_d) and one of A, B, or P,

or

(b) the projection matrix P and one of A or B.

In other words, if we assume $\mathfrak{D}$ is known and it can be interpreted as A, for example, we still do not have any information about the relative orientation of the term vectors or the dimensionality of the vector space. Thus, a critical question to be answered is with what concept from the physical problem do we want orthogonality or the degree of nonorthogonality (i.e., correlation) to be associated. Clearly, the choice we make at this level will directly influence the meaning that can be attached to document-query correlations, document-document correlations, and so on. When thought of in this framework, what we are looking for is a hypothesis that helps us determine the extent to which two terms are similar. Furthermore, it is quite likely that this information must be determined from outside, not implied in the term frequency data which we referred to as $\mathfrak{D}$.

Some earlier works in IR literature offer clues for how this might be done. Pseudo-classification [7] is a technique which has been the subject of several papers. The idea here is to obtain a classification of terms on the basis of relevance information obtained from the user. One of the papers in pseudo-classification [8] actually obtains a measure of relationship (correlation) between terms, rather than a classification of terms. Once we have a way to specify term-term correlations, it is then a simple matter to complete the picture vis-à-vis the vector space model.

Example 3. Consider a hypothetical collection which consists of two documents D_1 and D_2, each of which is described using 3 terms t_1, t_2, and t_3. That is, let

$$\mathfrak{D} = \begin{bmatrix} 2 & 3 \\ 3 & 7 \\ 5 & 1 \end{bmatrix}$$

Using the vector notation, we have

$$D_1 = 2t_1 + 3t_2 + 5t_3,$$

and

$$D_2 = 3t_1 + 7t_2 + t_3.$$

Clearly, elements of $\mathfrak{D}$ are interpreted as the components of D_is along t_is. Furthermore, suppose some technique such as pseudo-classification is used and the following term-term correlation matrix is obtained:

$$G_t = \begin{bmatrix} 1 & 0.5 & 0 \\ 0.5 & 1 & -0.3 \\ 0 & -0.3 & 1 \end{bmatrix}$$

Then, for a query $q_1 = (0, 0, 2)$

$$\begin{aligned} D_1 \cdot q_1 &= (2t_1 + 3t_2 + 5t_3) \cdot 2t_3 \\ &= 4t_1 \cdot t_3 + 6t_2 \cdot t_3 + 10t_3 \cdot t_3 \\ &= 4*0 - 6*0.3 + 10*1 \\ &= 8.2 \end{aligned}$$

For another query $q_2 = (2, 0, 0)$, the correlation is given by

$$\begin{aligned} D_1 \cdot q_2 &= 4t_1 \cdot t_1 + 6t_2 \cdot t_1 + 10t_3 \cdot t_1 \\ &= 4*1 + 6*0.5 + 10*0 \\ &= 7 \end{aligned}$$

Note that $D_1 \cdot q_1$ would have a higher value if the assumption that $t_2 \cdot t_3 = 0$ had been made. This is a direct result of the fact that $t_2 \cdot t_3$ is negative. The opposite effect is observed for $D_1 \cdot q_2$, since $t_2 \cdot t_1$ is positive.

It can be verified that G_t is nonsingular, indicating that all three terms are needed to span the subspace of interest. Just knowing D_1 and D_2 is not enough to have a basis consisting of documents alone (not enough of them). Furthermore, since G_t is nonsingular, there is no redundancy in the representation of document vectors D_1, D_2 (i.e., their expansions in terms of t_1, t_2, and t_3 are unique). □

An interesting alternative to the above is to adopt some scheme to obtain document-document correlations externally. For example, citation information that indicates the pattern in which documents refer to each other can be used to specify what we mean by correlation between two documents.

With either of these approaches, there would still be a need to expend the effort on the problem of determining the rank of the correlation matrix in order to identify a set of basis vectors. This step is required if the correlation matrix is singular (that is, its determinant equals zero). In that case, the rank of the matrix can be determined by identifying the largest principal submatrix whose determinant is nonzero. It is worth noting, however, that the above approaches really ask the question "How can the correlation between terms (or documents) be measured?" and the question of dimensionality is deferred. It is also possible to propose some other hypothesis which can directly resolve the question of dimensionality. The hypothesis would have to be more realistic than saying that the dimension is equal to the number of distinct terms. This we believe will help get around the problem of having to determine the rank of the correlation matrix. We have, in fact, developed a scheme taking this latter approach [9]. It would take too long to provide the details of the scheme here. Instead, we present a few highlights. It is assumed that two *concepts* that never jointly appear in any document are orthogonal. This means that atomic concepts, which are formed by conjunction of original terms such that each is negated or unnegated, are always pairwise orthogonal. Each original term is, then, expressed as a linear combination of vectors associated with certain atomic concepts. The coefficients used in the expansion will depend on the properties of documents which characterize the atomic concepts involved. The term-term correlations are then obtained by the scalar product of the various term vectors. The remaining steps are a matter of working with Eqs. (3) and (6) developed earlier.

Although this assumption would suggest that the number of dimensions is potentially exponential in the number of terms, it is necessary to be concerned only with the atomic concepts actually present in the document collection. Consequently, the number of distinct atomic concepts that appear in the collection is bounded by the number of documents. Some initial results using these ideas have been reported in [9].

Regardless of the specific approach used for this step, the vector space model prescribes a method for how correlations should be used when they are available. For example, if term-term correlations are known, then Eq. (4) shows exactly how that information should be incorporated for retrieval purposes. Earlier work on vector space model, surprisingly, has given no consideration to this apparently natural scheme. Examples of proposals that have appeared for this step are to construct clusters of terms (thesaurus) and then to incorporate this information for retrieval by substituting terms in documents and queries by term clusters (concepts) [10,11] or to expand the query with terms that belong to the same cluster as those in query [12].

C. A Case for Negative Components or Correlations

Before leaving this section, it is in order to reconsider the suggestion in the current use of the vector model, that the vector elements should be positive.

In the framework presented in this article, there is no

reason to prescribe that all the matrix elements of A, B, P, G_i, or G_{ii} are necessarily positive numbers. In fact, both negative and positive vector elements are appropriate and necessary as can be seen from the following example.

Example 4. In the two-dimensional vector space shown in Figure 3, the document vectors, D_1 and D_2, are expressed as a linear combination of the basis term vectors, t_1 and t_2:

$$D_1 = a_{11}t_1 + a_{21}t_2,$$

$$D_2 = a_{12}t_1 + a_{22}t_2.$$

It is easily seen from Figure 3 that if D_1 and D_2 are seen as the basis vectors, then the components of t_1 along the basis vectors are positive; so are the components of t_2 along the two documents. However, when the situation is turned around, and t_1, t_2 are considered to be the basis vectors, the component a_{21} of D_1 along t_2 and the component a_{12} of D_2 along t_1 are negative numbers. Although in this example all the projections are positive, it is also possible to have negative projections and/or correlations. Clearly, it is easy to imagine two index terms to be "opposite" in meaning and negative correlations would be a way to model that situation. In fact, the paper on pseudo-classification by one of the authors [8] discusses a scheme to determine both positive and negative relationship between terms. Based on the arguments presented here, the need to introduce negative components or correlations in the vector space model is quite evident.

6. Issues Relating to Linearly Dependent Set of Vectors

The standard vector space model assumes that terms are not correlated. More precisely, this mean that terms are pairwise orthogonal. Given some n terms, if they are pairwise orthogonal then it follows that the set of n vectors in question are also linearly independent. That is, the idea of pairwise orthogonality and the set being linearly independent coincide. But, if we have a set of n term vectors where certain pairs of vectors are correlated (not orthogonal), then it is not immediately clear whether the set is linearly independent. The two notions are, certainly, closely related since we need to know the exact pattern of correlations before we can determine the size of a maximal subset of vectors that are linearly independent. In other words, if $\{t_1, t_2, \ldots t_n\}$ represents the generating set for the subspace of interest and if the set is linearly dependent, then the representation of document vectors in terms of this set will not be unique. However, if we identify a (maximal) subset of vectors that are linearly independent, then the other vectors in the space have a unique representation in terms of this basis.

Thus, linear independence does not mean terms (or documents) are uncorrelated. Rather, linear independence only implies that any redundancy in the usage of terms has been removed and the representation in terms of the resulting set of vectors is compact (and unique). In the earlier literature, this separation of nonorthogonality and linear independence is not at all clear due to the fact that the assumption that terms are pairwise orthogonal was made to start with.

It is worth reiterating that the main objective of reduction of the correlation matrix is to determine the dimensionality of the space. Thus, as indicated in Section 5B, if some reasonable hypothesis can be made regarding dimensionality, the computational burden associated with the reduction process can be minimized.

The notions connected with reduction also share a close connection to earlier work in IR under the heading of discrimination value model [4,13,14]. After all, when we identify a linearly independent set from a linearly dependent set of term vectors, we are getting rid of unnecessary or superfluous terms and are thereby obtaining a space of reduced dimensions.

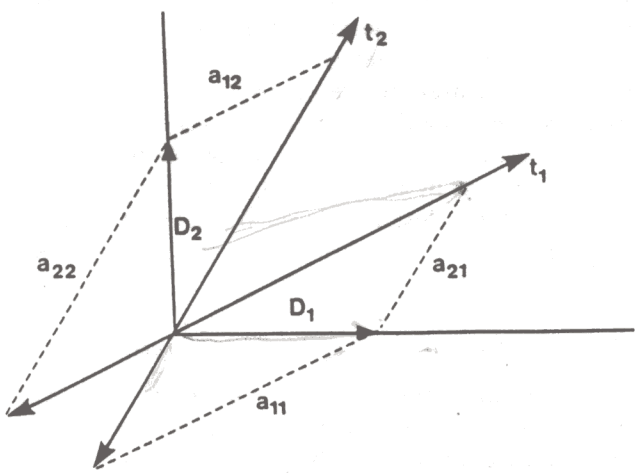

FIG. 3. Negative components in a two-dimensional vector space.

7. Conclusions

The way in which the vector space model has been introduced and used in the literature has led to a situation where many important concepts are ignored or poorly understood. In this work, we critically review the vector space model for IR by using notation that more clearly brings out problems and challenges associated with the use of the vector space model. In particular, the following assertions capsulate the main results of this investigation:

(i) Linear dependence of a set of (term) vectors, orthogonality or nonorthogonality of the various pairs of (term) vectors, and the dimensionality of the vector space are closely connected concepts.

(ii) Given a set of term vectors, every pair of vectors in the set being orthogonal implies that the set is linearly independent; the converse need not be true.

(iii) A vector space spanned by a linearly dependen

set of (term) vectors can always be spanned by a linearly independent subset of these vectors; thus, representation of a vector using a linearly dependent set of (term) vectors may consist of terms that can be eliminated without any loss of information.

(iv) Given the term-document matrix, consisting of frequencies of occurrence of terms in documents, they need not always be interpreted (as has been commonly done) as the components of document vectors along the terms; a number of other interpretations are possible.

(v) Given the term-document matrix and the interpretation of them as components of document vectors along terms, the model is not complete. One way to make this complete is to have a way to determine term-term similarities.

(vi) Without additional assumptions, computation of a quantity such as term-term similarities must be done, from outside, independently of term-document matrix. In particular, interpretation of the elements of the term-document matrix as both the component of documents along term vectors as well as the components of terms along document vectors (which has often been done) is inconsistent. That is, the choice of one makes the choice of the other incorrect.

We believe that this work will lead to the harnessing of the real power inherent in the vector space model as a formal framework for developing information retrieval systems.

References

1. Robertson, S. E.; Maron, M. E.; Cooper, W. S. "Probability of Relevance: A Unification of Two Competing Models for Document Retrieval." *Information Technology: Research and Development*. 1(1):1-21; 1982.
2. Salton, G.; McGill, M. H. *Introduction to Modern Information Retrieval*. New York: McGraw-Hill; 1983.
3. Salton, G. (ED.). *The SMART Retrieval System-Experiments in Automatic Document Processing*. Englewood Cliffs, NJ: Prentice-Hall; 1971.
4. Salton, G. *Dynamic Library and Information Processing*. Englewood Cliffs, NJ: Prentice-Hall; 1975.
5. Koll, M. "An Approach to Concept Based Information Retrieval." *ACM-SIGIR Forum*. XIII:32-50; 1979.
6. Greub, W. H. *Linear Algebra*. New York: Academic; 1963.
7. Salton, G. "Automatic Term Class Construction Using Relevance—A summary of Work in Automatic Pseudoclassification." *Information Processing and Management*. 16(1):1-15; 1980.
8. Raghavan, V. V.; Yu, C. T. "Experiments on the Determination of the Relationships Between Terms." *ACM Trans. on Database Systems*. 4(2):240-260; 1979.
9. Wong, S. K. M.; Ziarko, W.; Wong, P. C. N.. "Generalized Vector Space Model in Information Retrieval II." *Proc. of ACM-SIGIR Conference on Research and Development in Information Retrieval*. June 1985, Montreal, Canada.
10. Salton, G. "Experiments in Automatic Thesauras Construction for Information Retrieval." *Information Processing 71*. Amsterdam, The Netherlands: North-Holland; 1972:115-123.
11. Sparck-Jones, K. *Automatic Keyword Classifications*. London: Butterworths; 1971.
12. Minker, J.; Wilson, G. A.; Zimmerman, B. H. "An Evaluation of Query Expansion by the Addition of Clustered Terms for a Document Retrieval System." *Info. Stor. and Retrieval*. 8:329-348; 1972.
13. Can, F.; Ozkarahan, E. A. "Concepts of the Cover-Coefficient-Based Clustering Methodology." *Proc. of ACM-SIGIR Conference on Research and Development in Information Retrieval*. June 1985, Montreal, Canada.
14. Can, F.; Ozkarahan, E. A. "Similarity and Stability Analysis of the Two Partitioning Type Clustering Algorithms." *Journal of the American Society of Information Science*. 36(1):3-14; 1985.