Generalized Vector Space Model In Information Retrieval

S.K.M. Wong, Wojciech Ziarko and Patrick C.N. Wong
Department of Computer Science, University of Regina,
Regina, Sask., Canada S4S 0A2

Abstract. In information retrieval, it is common to model index terms and documents as vectors in a suitably defined vector space. The main difficulty with this approach is that the explicit representation of term vectors is not known a priori. For this reason, the vector space model adopted by Salton for the SMART system treats the terms as a set of orthogonal vectors. In such a model it is often necessary to adopt a separate, corrective procedure to take into account the correlations between terms. In this paper, we propose a systematic method (the generalized vector space model) to compute term correlations directly from automatic indexing scheme. We also demonstrate how such correlations can be included with minimal modification in the existing vector based information retrieval systems. The preliminary experimental results obtained from the new model are very encouraging.

1. Introduction

In the vector space model proposed by Salton [1,2,3], the keywords or index terms are viewed as basic vectors in a linear vector space, and each document is represented as a vector in such a space. It can be argued that the frequency of occurrence of a term in a document represents the component of the document along the corresponding basic term vector. However, if only the occurrence frequency for each term is available, it is not possible to characterize the vector space completely [11]. Either we need to know the explicit representation of the term vectors or we need some assumptions to account for the correlations between terms. For instance, in the SMART system the term vectors are assumed to be orthogonal. Since terms are, in fact, correlated, it is often necessary in such an approach to introduce a separate, corrective measure for incorporating term correlations in some ad hoc fashion.

One well known method for computing term correlations is based on term co-occurrence frequencies. However, the use of a co-occurrence matrix can be justified only if the documents and term vectors are assumed to be orthogonal. Several authors have proposed different methods of recognizing term correlations in the retrieval process. Raghavan and Yu [4] used a statistical analysis of queries vs. relevant and nonrelevant documents in order to determine positive and negative correlations among terms. A probablistic approach to the problem of term dependency was presented by Van Rijsbergen and Harper [5,6]. Their basic assumption is that index terms are distributed in a dependent manner in the document space. However, the resulting formula for computing the dependency factors does not seem computationally feasible even for a relatively small number of terms [7]. Katter [8] and Switzer [9] started from a term co-occurrence matrix and derived a basic set of term vectors through techniques of factor analysis or multi-dimensional scaling. This approach has the advantage that the terms are not treated as though they are linearly independent. Recently, Koll [10], on the other hand, developed a scheme by which correlations between terms can be incorporated without having to handle the term co-occurrence matrix. The difficulty with this latter approach is that it does not have an adequate formal justification.

We believe up to the present time that there is no satisfactory way of computing term correlations based on automatic indexing scheme. The current work has objectives similar to the studies mentioned above. We propose a new method to represent term vectors explicitly in terms of a suitably chosen set of orthonormal basic vectors. This means that term correlations can then be computed directly from such a representation. In contrast to many recent studies, it is not necessary in our approach to assume that either the document or the term vectors have to be orthogonal. We also demonstrate how such term correlations can be included in a natural manner in the existing vector based information retrieval systems (e.g. in the SMART system) with minimal modifications.

Before the basic model (hereafter referred to as the generalized vector space model or GVSM) is introduced in Section 4, we will first use two simple examples to illustrate how term correlations can be computed from an intuitive point of view. In Section

6 we select two standard collections of documents to evaluate the retrieval performance of the GVSM in comparison with the conventional vector space model (VSM).

2. Basic Definitions And Concepts In The Conventional Vector Space Model (VSM)

The basic premise in the vector space model is that the documents and the query are represented by a set of vectors, say, $\{\vec{d}_\alpha\}$, $\alpha = 1, 2, \ldots, p$, and $\vec{q}$, respectively, in a vector space spanned by the normalized term vectors, $\{\vec{t}_i\}$, $i = 1, 2, \ldots, n$. That is,

$$\vec{d}_\alpha = \sum_{i=1}^{n} a_{\alpha i} \vec{t}_i , \qquad (\alpha = 1, 2, \ldots, p), \quad (1a)$$

$$\vec{q} = \sum_{j=1}^{n} q_j \vec{t}_j . \qquad (2)$$

Given the above representation for $\vec{d}$ and $\vec{q}$, for example, the scalar product $\vec{d}_\alpha \cdot \vec{q}$, which may serve as a measure of the similarity between each document in $\{\vec{d}_\alpha\}_p$ and the query $\vec{q}$, is defined by

$$\vec{d}_\alpha \cdot \vec{q} = \sum_{i=1, j=1}^{n} a_{\alpha i} q_j \vec{t}_i \cdot \vec{t}_j , \quad (\alpha = 1, 2, \ldots, p) . \quad (3a)$$

We can, then, rank the documents with respect to the query $\vec{q}$ according to the values of the above similarity function. Thus, for our purpose it is necessary to know both the correlations between the vectors, $\{\vec{t}_i\}_n$, and the components of documents and queries along these basic vectors.

It is convenient in subsequent discussions to express equation (1a) in matrix notation as follows:

$$\vec{D} = \vec{t}\, A^T , \qquad (1b)$$

where

$\vec{D} = (\vec{d}_1, \vec{d}_2, \ldots, \vec{d}_p)$, $\vec{t} = (\vec{t}_1, \vec{t}_2, \ldots, \vec{t}_n)$, and

$$A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ a_{p1} & a_{p2} & \cdots & a_{pn} \end{bmatrix} \qquad (4)$$

Similarly, equation (3a) can be rewritten as

$$\vec{S} = \vec{q}\, G\, A^T , \qquad (3b)$$

where $\vec{S} = (\vec{d}_1 \cdot \vec{q}, \vec{d}_2 \cdot \vec{q}, \ldots, \vec{d}_p \cdot \vec{q})$,

$\vec{q} = (q_1, q_2, \ldots, q_n)$, and

$$G = \begin{bmatrix} \vec{t}_1 \cdot \vec{t}_1 & \vec{t}_1 \cdot \vec{t}_2 & \cdots & \vec{t}_1 \cdot \vec{t}_n \\ \vec{t}_2 \cdot \vec{t}_1 & \vec{t}_2 \cdot \vec{t}_2 & \cdots & \vec{t}_2 \cdot \vec{t}_n \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \vec{t}_n \cdot \vec{t}_1 & \vec{t}_n \cdot \vec{t}_2 & \cdots & \vec{t}_n \cdot \vec{t}_n \end{bmatrix} \qquad (5)$$

In the conventional vector space model, the matrix A is assumed to be the term occurrence frequency matrix empirically obtained from automatic indexing. Since correlations between terms are not known a priori, as a first order of approximation, the correlation matrix G defined by equation (5) is assumed to be an identity matrix. With such approximations (i.e. $G = I$), the ranking vector $\vec{S}$ for a given query $\vec{q}$ can, therefore, be computed easily from the following equation:

$$\vec{S} = \vec{q}\, A^T . \qquad (6)$$

The strength of such an approach clearly lies in its simplicity. However, one of its main drawbacks is that it ignores term correlations. Very often, one has to modify the above similarity function (6) by introducing some ad hoc schemes for including the important effect of term correlations. In Section 4, we suggest a method to compute term correlations by representing the term vectors explicitly in a vector space spanned by the atoms of a free boolean algebra generated by the index terms. Consequently, term correlations can be incorporated directly through equation (3b) in order to obtain higher retrieval performance without the need to modify the similarity function or to introduce a new one.

3. Term Correlations

Before developing our model formally in the next section, it is fitting, perhaps, to demonstrate first how term correlations can be computed from an intuitive point of view. Let us consider two simple examples.

Example 1. Let D be a set of documents indexed only by two terms, t_1 and t_2.

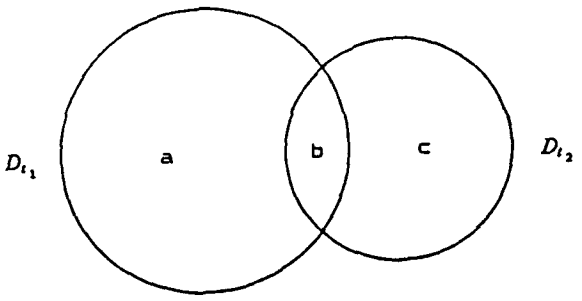

Figure 1. Partition of D into disjoint subsets a, b, and c.

In Figure 1, the subsets a, b and c of D are defined by :

$$a = D_{t_1\bar{t}_2} = D_{t_1} \cap \bar{D}_{t_2},$$

$$b = D_{t_1t_2} = D_{t_1} \cap D_{t_2},$$

$$c = D_{\bar{t}_1t_2} = \bar{D}_{t_1} \cap D_{t_2},$$

where D_{t_i} , (i = 1, 2) , is the maximal subset of D containing t_i , and $\bar{D}_{t_i}$ denotes the set complement of D_{t_i} (i.e. $\bar{D}_{t_i}$ is the subset of documents not containing t_i) .

Based on intuition we argue that the correlation between any two index terms depends on the number of documents in which these two terms appear together. Let $c(D)$ denote the cardinality of an arbitrary set D. In Figure 1 the cardinality $c(D_{t_1t_2})$ of the subset $b = D_{t_1t_2} = D_{t_1} \cap D_{t_2}$ (which denotes the number of documents containing t_1 and t_2) thus provides a plausible measure of the "unnormalized" correlation between t_1 and t_2 .

In terms of vector notation , the normalized correlation between t_1 , and t_2 , denoted by $\vec{t}_1 \cdot \vec{t}_2$, can be conveniently expressed as the scalar product of two normalized term vectors, $\vec{t}_1$ and $\vec{t}_2$, namely ,

$$\vec{t}_1 \cdot \vec{t}_2 = \frac{c^2(D_{t_1t_2})}{[c^2(D_{t_1\bar{t}_2}) + c^2(D_{t_1t_2})]^{1/2}\,[c^2(D_{\bar{t}_1t_2}) + c^2(D_{t_1t_2})]^{1/2}},$$

where

$$\vec{t}_1 = \frac{c(D_{t_1\bar{t}_2})\,\vec{m}_1 + c(D_{t_1t_2})\,\vec{m}_2}{[c^2(D_{t_1\bar{t}_2}) + c^2(D_{t_1t_2})]^{1/2}},$$

$$\vec{t}_2 = \frac{c(D_{t_1t_2})\,\vec{m}_2 + c(D_{\bar{t}_1t_2})\,\vec{m}_3}{[c^2(D_{\bar{t}_1t_2}) + c^2(D_{t_1t_2})]^{1/2}},$$

and $\vec{m}_1$, $\vec{m}_2$, and $\vec{m}_3$ are some suitably chosen orthonormal basic vectors .

Example 2. The main purpose of this example is to show that the concepts introduced in Example 1 can be easily generalized in a more complicated situation. Consider the partition of a set of documents D indexed by terms t_1 , t_2 , and t_3 as shown in Figure 2.

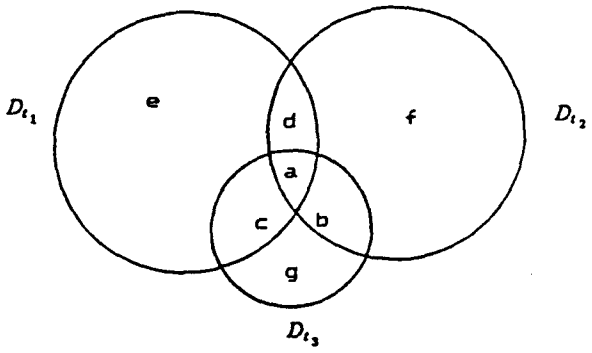

Figure 2. Partition of D into disjoint subsets a, b , c, d, e, f and g.

The disjoint subsets a , b , c , e , f , and g of D can be specified as follows :

$$a = D_{t_1t_2t_3} = D_{t_1} \cap D_{t_2} \cap D_{t_3},$$

$$b = D_{\bar{t}_1t_2t_3} = \bar{D}_{t_1} \cap D_{t_2} \cap D_{t_3},$$

$$c = D_{t_1\bar{t}_2t_3} = D_{t_1} \cap \bar{D}_{t_2} \cap D_{t_3},$$

$$d = D_{t_1t_2\bar{t}_3} = D_{t_1} \cap D_{t_2} \cap \bar{D}_{t_3},$$

$$e = D_{t_1\bar{t}_2\bar{t}_3} = D_{t_1} \cap \bar{D}_{t_2} \cap \bar{D}_{t_3},$$

$$f = D_{\bar{t}_1t_2\bar{t}_3} = \bar{D}_{t_1} \cap D_{t_2} \cap \bar{D}_{t_3},$$

$$g = D_{\bar{t}_1\bar{t}_2t_3} = \bar{D}_{t_1} \cap \bar{D}_{t_2} \cap D_{t_3},$$

where D_{t_i} , (i = 1 , 2 , 3) , is the maximal subset of D containing t_i .

As in Example 1, term correlations $\vec{t}_i \cdot \vec{t}_j$ for $1 \leq i < j \leq 3$ can be intuitively expressed as the scalar products of the following normalized term vectors $\vec{t}_1$, $\vec{t}_2$, and $\vec{t}_3$:

$$\vec{t}_1 \cdot \vec{t}_2 = \frac{c^2(D_{t_1t_2t_3}) + c^2(D_{t_1t_2\bar{t}_3})}{N_1 N_2},$$

$$\vec{t}_1 \cdot \vec{t}_3 = \frac{c^2(D_{t_1 t_2 t_3}) + c^2(D_{t_1 \bar{t}_2 t_3})}{N_1 N_3},$$

$$\vec{t}_2 \cdot \vec{t}_3 = \frac{c^2(D_{t_1 t_2 t_3}) + c^2(D_{\bar{t}_1 t_2 t_3})}{N_2 N_3},$$

where

$$\vec{t}_1 = [c(D_{t_1 t_2 t_3})\vec{m}_1 + c(D_{t_1 \bar{t}_2 t_3})\vec{m}_3] \;/\; N_1 + [c(D_{t_1 t_2 \bar{t}_3})\vec{m}_4 + c(D_{t_1 \bar{t}_2 \bar{t}_3})\vec{m}_5] \;/\; N_1,$$

$$\vec{t}_2 = [c(D_{t_1 t_2 t_3})\vec{m}_1 + c(D_{\bar{t}_1 t_2 t_3})\vec{m}_2] \;/\; N_2 + [c(D_{t_1 t_2 \bar{t}_3})\vec{m}_4 + c(D_{\bar{t}_1 t_2 \bar{t}_3})\vec{m}_6] \;/\; N_2,$$

$$\vec{t}_3 = [c(D_{t_1 t_2 t_3})\vec{m}_1 + c(D_{\bar{t}_1 t_2 t_3})\vec{m}_2] \;/\; N_3 + [c(D_{t_1 \bar{t}_2 t_3})\vec{m}_3 + c(D_{\bar{t}_1 \bar{t}_2 t_3})\vec{m}_7] \;/\; N_3,$$

$$N_1^2 = c^2(D_{t_1 t_2 t_3}) + c^2(D_{t_1 \bar{t}_2 t_3}) + c^2(D_{t_1 t_2 \bar{t}_3}) + c^2(D_{t_1 \bar{t}_2 \bar{t}_3}),$$

$$N_2^2 = c^2(D_{t_1 t_2 t_3}) + c^2(D_{\bar{t}_1 t_2 t_3}) + c^2(D_{t_1 t_2 \bar{t}_3}) + c^2(D_{\bar{t}_1 t_2 \bar{t}_3}),$$

$$N_3^2 = c^2(D_{t_1 t_2 t_3}) + c^2(D_{\bar{t}_1 t_2 t_3}) + c^2(D_{t_1 \bar{t}_2 t_3}) + c^2(D_{\bar{t}_1 \bar{t}_2 t_3}),$$

and $\vec{m}_1, \vec{m}_2, \vec{m}_3, \vec{m}_4, \vec{m}_5, \vec{m}_6$, and $\vec{m}_7$ are orthonormal basic vectors.

We will show that the above results can be formally derived as a special case from the generalized vector space model developed in the following section.

4. The Generalized Vector Space Model (GVSM)

As we mentioned in Section 2 there is no explicit knowledge about the term vectors nor the vector space itself in the conventional vector space model (VSM). We will show, in contrast to the VSM, that the representation of term vectors can be explicitly defined in a 2^n - dimensional cartesian space. This requires the introduction of the notion of a boolean algebra. (It is assumed, in the following discussion, that the reader is already familiar with the representation of a boolean algebra and its related concepts).

4.1. Vector Representation of a Free Boolean Algebra (without External Constraints)

We will first discuss the free boolean algebra B_{2^n} generated by n literals (index terms). A fundamental products in n literals, $t_1, t_2, \ldots t_n$ is a conjunction in which each literal t_j appears exactly once, either complemented or uncomplemented, for j = 1, 2, . . . , n. For example, $t_1 \bar{t}_2 t_3$ and $\bar{t}_1 t_2 \bar{t}_3$ are fundamental products in three literals, t_1, t_2, and t_3. In fact, the 2^n fundamental products in n literals are join-irreducible, and, therefore, they are minterms (atoms) of the free boolean algebra B_{2^n} containing 2^{2^n} elements. It follows that any boolean expression $E(t_1, t_2, \ldots, t_n)$, i.e. any element in B_{2^n}, can be transformed into a unique disjunctive canonical expression (sum of minterms). Let $\{m_k\}_{2^n}$ denote the set of minterms in B_{2^n}. Each minterm m_k can be labelled by $k = (\delta_1, \delta_2, \ldots, \delta_n)$, where $\delta_i = 0$ or 1 for $1 \leqslant i \leqslant n$. That is,

$$m_k = t_1^{\delta_1} \wedge t_2^{\delta_2} \wedge \ldots \wedge t_n^{\delta_n}, \qquad (7)$$

$$\text{where } t_i^{\delta_i} = \begin{cases} t_i & \text{if } \delta_i = 1, \\ \bar{t}_i & \text{if } \delta_i = 0. \end{cases}$$

In particular, since each literal t_i is itself an element in B_{2^n}, every t_i, for $1 \leqslant i \leqslant n$, can be expressed uniquely in terms of the disjunction of minterms, namely,

$$t_i = m_{i_1} \vee m_{i_2} \vee \ldots \vee m_{i_j} \vee \ldots \vee m_{i_r}, \qquad (8)$$

for all those m_{i_j}'s such that $m_{i_j} \vee t_i = t_i$.

We can, therefore, represent the elements of the free boolean algebra B_{2^n}, as a unitary 2^n - cube in a 2^n - dimensional cartesian space, R^{2^n}. In terms of vector notation, the set of minterms, $\{m\}_{2^n}$, in B_{2^n} can be represented explicitly as the set of orthonormal basic vectors, $\{m\}_{2^n}$, in R^{2^n} as follows:

$$\begin{aligned} \vec{m}_1 &= (1,0,0,\ldots,0), \\ \vec{m}_2 &= (0,1,0,\ldots,0), \\ \vec{m}_3 &= (0,0,1,\ldots,0), \\ &\vdots \\ \vec{m}_{2^n} &= (0,0,0,\ldots,1). \end{aligned} \qquad (9)$$

Hence, in the absence of any external constraints, each literal t_i, defined by equation (8), has the following unique vector representation:

$$\vec{t}_i = \sum_{j=1}^{r} \vec{m}_{i_j}.$$

For example, the elements of B_3 can be represented as a 3 - dimensional cube as shown in Figure 3, where $\vec{m}_1 = (1,0,0)$, $\vec{m}_2 = (0,1,0)$, and $\vec{m}_3 = (0,0,1)$ are the orthonormal basic vectors of R^3.

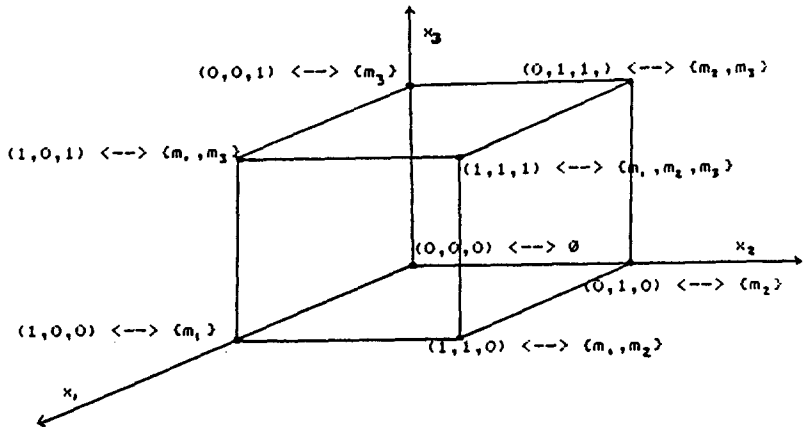

Figure 3. The 3 - Dimensional Cube of B_3.

4.2. Vector Representation of Index Terms in the GVSM

The mapping introduced in the previous section is applicable only to a strictly boolean retrieval system in which each document is represented by a vector as follows :

$$\vec{d}_\alpha = (a_{\alpha 1}, a_{\alpha 2}, \ldots, a_{\alpha n}), \qquad (10)$$

where

$$a_{\alpha i} = \begin{cases} 1 & \text{if } \textit{document } d_\alpha \textit{ is indexed by term } t_i \\ 0 & \text{if } \textit{document } d_\alpha \textit{ is not indexed by term } t_i . \end{cases}$$

In a more realistic information retrieval model, relative term weights ($0 \leqslant a_{\alpha i} \leqslant 1$) are assigned to each document d_α . Our main objective here is to define a more general mapping which enables us to transform each index term into a vector in R^{2^n} for a collection of documents, D = {d}, with term weights other than 0 or 1. For this purpose, we introduce a composite transformation $gf : \{t\}_n \cup \{m\}_{2^n} \rightarrow R^{2^n}$, where the individual mappings f and g are defined as follows :

(i) $f : \{t\}_n \cup \{m\}_{2^n} \rightarrow 2^{\{d\}}$ ($2^{\{d\}}$ denotes the power set of D = {d})

(a) For any index term $t_i \in \{t\}_n$, $f(t_i)$ is the maximal subset D_{t_i} of documents each of which contains term t_i .

(b) For any atom $m_k = t_1^{\delta_1} \wedge t_2^{\delta_2} \wedge \ldots \wedge t_n^{\delta_n}$ as defined by equation (7),

$$f(m_k) = f(t_1^{\delta 1}) \cap f(t_2^{\delta 2}) \cap \ldots \cap f(t_n^{\delta n}), \qquad (11)$$

where

$$f(t_i^{\delta_i}) = \begin{cases} f(t_i) = D_{t_i} & \text{if } \delta_i = 1 , \\ f(\bar{t}_i) = \bar{D}_{t_i} & \text{if } \delta_i = 0 , \end{cases} \qquad (12)$$

and $\bar{D}_{t_i}$ denotes the set complement of D_{t_i} .

(ii) $g : f(\{t\}_n \cup \{m\}_{2^n}) \rightarrow R^{2^n}$

(a) For any atom m_k in the boolean algebra B_{2^n} ,

$$g(f(m_k)) = \vec{m}_k , \qquad (13)$$

where $\vec{m}_k$ is the k-th orthonormal basic vector of R^{2^n} , which is defined by equation (9).

(b) From equations (8) and (13), for any index term $t_i \in \{t\}_n$, the term vector $\vec{t}_i$ is defined by the following transformation :

$$\vec{t}_i = g(f(t_i)) = \sum_{k=1}^{r} c_k(t_i)\, g(f(m_{i_k})) = \sum_{k=1}^{r} c_k(t_i)\vec{m}_{i_k} , \qquad (14)$$

where

$$c_k(t_i) = \sum_{\alpha \in I(t_i, k)} a_{\alpha i} , \qquad (15)$$

and the above summation is restricted to the set:

$$I(t_i, k) = \{\alpha \mid d_\alpha \in f(m_{i_k}) \subseteq f(t_i)\} .$$

It is important to note that in equation (15) the $a_{\alpha i}$'s are the matrix elements of the term occurrence frequency matrix A defined by equation (4). We have, therefore, shown that by the above composite mapping fg each index term t_i can be transformed into a vector $\vec{t_i} = g(f(t_i))$ in R^{2^n} based on the input term occurrence frequency matrix [7].

It follows immediately from equations (14) and (15) that the normalized term vectors can be expressed as

$$\vec{t_i} = \frac{1}{N_i} \sum_{k=1}^{r} c_k(t_i) \vec{m}_{i_k} , \quad 1 \leqslant i \leqslant n , \qquad (16)$$

where

$$N_i^2 = \sum_{k=1}^{r} c_k^2(t_i) . \qquad (17)$$

The term correlation, $\vec{t_i} \cdot \vec{t_j}$, between any pair of index terms, t_i and t_j, can now be computed directly from equation (16).

If we substitute the $a_{\alpha i}$'s given by equation (10) into equation (15), we obtain

$$c_k(t_i) = c(f(m_{i_k})) , \qquad (18)$$

where $c(f(m_{i_k}))$ is the cardinality of the subset of documents defined by the atom m_{i_k} in equation (11). By substituting the values for $c_k(t_i)$ given by equation (18) into equation (16), we immediately arrive at the same expansions for the term vectors as those given in Section 3. It is clear that the results we obtained informally in Section 3 are special cases which can be easily derived from the generalized vector space model presented here.

Similar to the conventional vector space model, we may assume that the document space can be spanned by the normalized term vectors as in equation (1a). It should be emphasized that the term vectors, $\{\vec{t_i}\}_n$, are now explicitly defined by equation(16). We can easily rank the retrieval outputs in decreasing order of the query-document similarities (which are computed directly from equation(3a)). The main advantage of our method is that not only are term correlations explicitly known, but also the effect of these correlations can be naturally incorporated through equation (3a) without the need to introduce any ad hoc similarity function as in other vector based information retrieval models.

5. An Algorithm For Spanning The Term Vectors

In this section we would like to suggest an algorithm for computing term correlations based on the generalized vector space model developed in Section 4. This algorithm is not necessarily an optimal method but it serves the purpose for demonstrating how the concepts we have introduced can be applied. The input to this algorithm is simply the term occurrence frequency matrix A obtained from automatic indexing and the output is term correlation between any pair of index terms.

Step (i) For a given term t_i, identify the maximal subset of documents, D_{t_i}, containing t_i by the set of non-zero matrix elements in the i-th column of A. Construct a submatrix $A(t_i)$ by deleting all the rows of A with zero elements in the i-th column.

Step (ii) Partition the rows of $A(t_i)$ into r blocks each of which corresponds to a distinct subset of index terms. Each of these subsets of terms, in fact, represents an atom m_{i_k} which has a non-zero coefficient $c_k(t_i)$ in the expansion for vector $\vec{t_i}$ (see equation (14)). From equation (15), compute the sum $c_k(t_i) = \sum_{\alpha} a_{\alpha i}$ for each k-th block. The normalization factor N_i^2 defined by equation (17) can be calculated by summing $c_k^2(t_i)$ for k = 1, 2, . . ., r. Repeat steps (i) and (ii) to obtain the expansions for other term vectors.

6. Experimental Evaluation of The GVSM

For the preliminary evaluation of the generalized vector space model, we use two collections of documents, ADINUL(82 documents and 35 queries) and CRN4NUL(424 documents and 155 queries), which are standard test data in the conventional VSM . These collections include, for evaluation purposes, information for each query as to which of the documents are relevant. The standard recall and precision measures are used for comparing the retrieval performance of the GVSM and VSM in the above two document collections. Recall is defined as the proportion of relevant documents retrieved, while precision is the proportion of retrieved documents that are relevant [3] . The overall performance of a retrieval strategy is determined by computing the average precision over all queries for recall values 0.1, 0.2, . . ., 1.0 . The procedure for averaging is consistent with that implemented in the SMART system.

The input to the GVSM is the term occurrence frequency matrix obtained from automatic indexing. In the first step of our computation, using equations (14), (15), and (16) , each normalized term vector t_i is expressed as a linear combination of the orthonormal basic vectors, $\{\vec{m}\}_{2^n}$. Term correlations are explicitly included through equation (3b) in ranking the retrieval outputs. The solid curves shown in Figures 4 and 5 represent the average recall-precision values

computed from the generalized vector space model. For the purpose of comparison, we also reproduce the query-document similarities based on equation (6) (the well known cosine similarity function) in which term correlations are ignored as in the conventional VSM. These results are summarized in the average recall-precision graphs as shown by the dotted lines in Figures 4 and 5.

Comparing the results for the GVSM to those of the VSM, it is seen that from the above analysis the former is significantly better in both collections of documents. We are planning for more extensive evaluation of the GVSM. Nevertheless, these preliminary results indicate the potential of the present approach both from the practical and theoretical points of view.

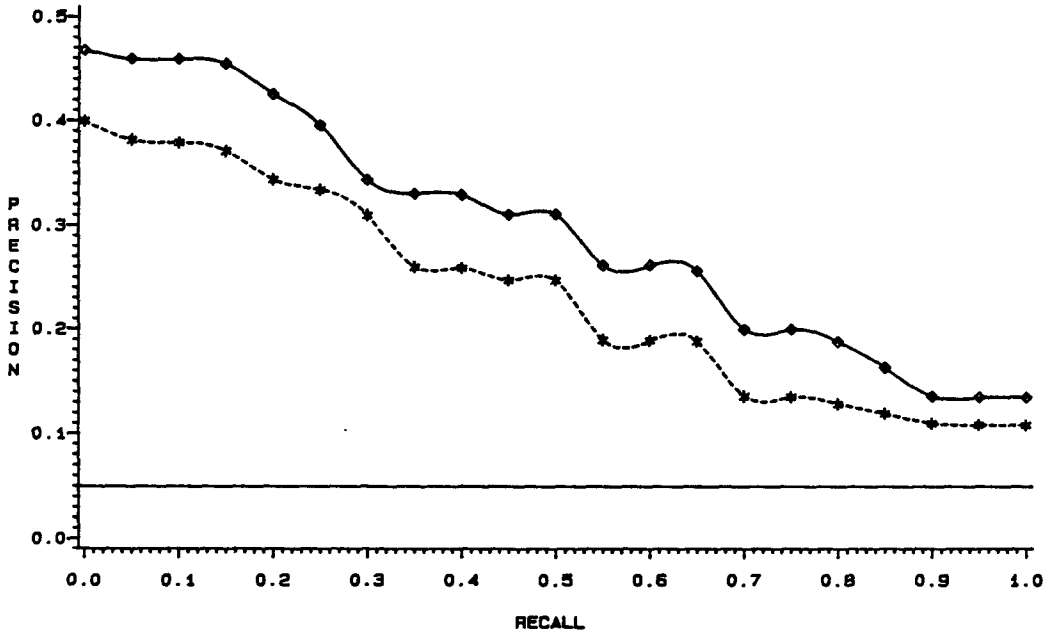

Figure 4. Comparsion of recall-precision between GVSM (solid curve) and VSM (dotted curve) in ADINUL.

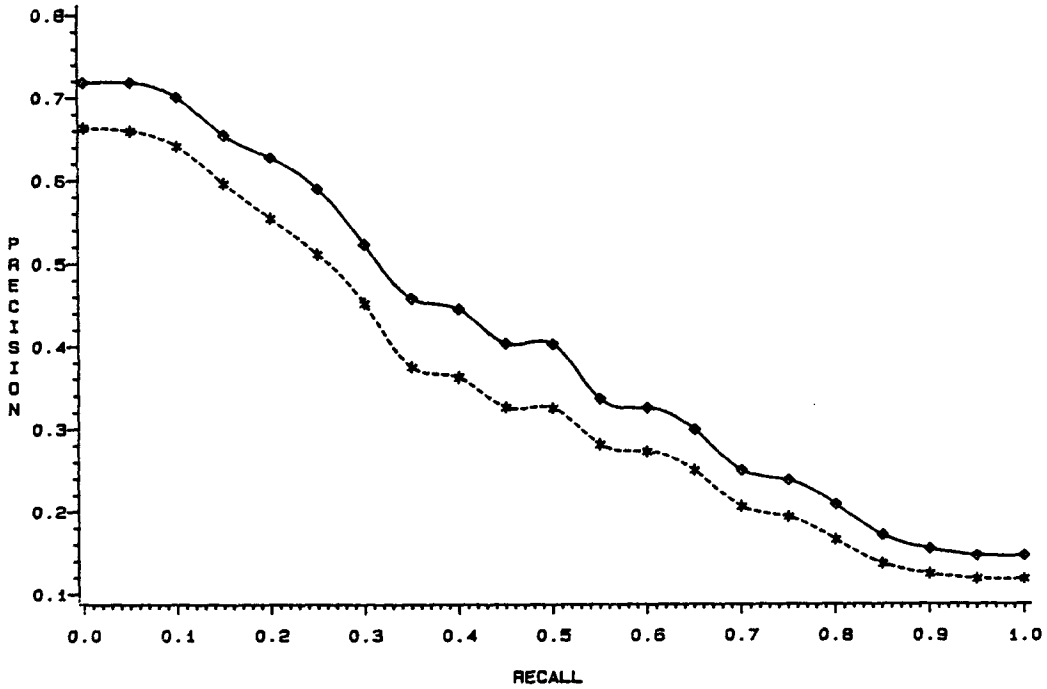

Figure 5. Comparsion of recall-precision between GVSM (solid curve) and VSM (dotted curve) in CRN4NUL.

7. Conclusion

The conventional vector space model has been proven in practice to be a valuable tool in information retrieval systems. Its main drawback lies in the lack of systematic methods for computing term correlations which we believe are crucial in the retrieval process. Although there are many recent papers which have similar objectives as ours, they lack, in our opinion, the theoretical foundations.

We have shown that term vectors can be explicitly represented in a 2^n - dimensional vector space based on the notion of boolean algebra. Another advantage of our method is that these term correlations can be incorporated into the existing vector-processing systems in a straightforward manner. We believe that many important concepts in the vector space model can be formally established and better understood within the framework of our approach.

At the present time, we are extending the capability of the GVSM to process boolean-like queries. We will report on the results of this investigation in the near future.

8. Acknowledgement

We are very grateful to Dr. Zdzislaw Pawlak, Institute of Computer Science, Polish Academy of Science, and our colleague Dr. V. V. Raghavan for their encouragement and valuable comments.

References

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

[2] Salton, G., Dynamic Information and Library Processing. Prentice-Hall, Englewood Cliffs, New Jersey. (1983).

[3] Salton, G. and McGill, M.J., Introduction to Modern Information Retrieval. McGraw Hill, New York. (1983).

[4] Raghavan, V.V. and Yu, C.T., Experiments on the Determination of the Relationships Between Terms. ACM Transactions on Database Systems no. 4, (1979). pp. 240 - 260.

[5] 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.

[6] Harper, D.J. and Van Rijsbergen, C.J., An Evaluation of Feedback in Document Retrieval using Co-occurrence Data. Journal of Documentation, vol 34, (1978). pp. 189 - 216.

[7] Salton, G., Buckley, C. and Yu, C.T., An Evaluation of Term Dependence Models in Information Retrieval. Proceedings of the 5-th ACM SIGIR Conference (1982).

[8] Katter, R.V., A study of Document Representations: Multidimension Scaling of Index Terms. SDC - Final Report, (1967).

[9] Switzer, P., Vector Images in Information Retrieval. Proceedings of the Symposium on Statistical Association Methods for Mechanical Documentation, Wash. D.C., 1964. (NBS Misc. Publ. 269, 1965) Stevens, M.E., Heilprin, L., Guiliano, V.E. (eds.). pp. 163 - 171.

[10] Koll, M., Weird: An Approach to Concept - based Information Retrieval. ACM - SIGIR Forum, vol XIII, no. 4, (spring 1979), pp. 32 - 50.

[11] Wong, S.K.M. and Raghavan, V.V., Vector Space Model of Information Retrieval - A Re-Evaluation. Proceedings of the 3-rd Joint BCS and ACM Symposium on Research and Development in Information Retrieval, Kings College, Cambridge, England. (1984).