Paradox-free Formal Foundation of Vector Space Model

Sándor Dominich

Department of Computer Science, University of Veszprém,
Egyetem u. 10, 8200 Veszprém, Hungary,
Email: dominich@dcs.vein.hu

Abstract

In the Vector Space Model (VSM) of Information Retrieval (IR), the documents and queries are conceived as vectors of weights belonging to the same linear space, and retrieval is based on similarity measures. This view yields paradoxes and problems as regards concepts of fundamental importance to IR (similarity, term independence, preference, linearity). In the present paper, a different view is proposed: the VSM of IR may be conceived as a probability space that decreases the amount of the Shannon information associated to the selection of documents as answers to queries. This view can serve as another formal foundation of the VSM that is paradox-free and, at the same time, preserves the computational practice developed thus far. Experiments are also reported to reflect this view in practice.

Key Words

Information Retrieval, Vector Space Model, Probability Space,

1 INTRODUCTION

Consider an entity described by a piece of text (traditionally called a document). If its words are ranked in decreasing order with respect to their number of occurrences (also called frequencies), the product between the rank of any word and its umber of occurrences is approximately constant (Zipf, 1949). If it is assumed, naturally enough, that the most obvious place where appropriate content identifiers might be found is the document itself, then the number of occurrences of its terms can give meaningful indication of its content (Luhn, 1959). Consider now *m* documents and *n* terms. Each document can be assigned a sequence (of length *n*) of weights, which represent the degrees to which the terms pertain to (characterise) that document. If all these sequences are put together an $n \times m$ matrix, called term-by-document matrix, of weights is obtained, where the columns correspond to documents, while the rows to terms. Let us consider a textual query expressing an information need to which an answer is to be found by searching the documents. In (Salton, 1965), it was proposed that both the documents and queries should use the same conceptual space, while in (Salton et al., 1975) this idea was combined with the term-by-document matrix. Thus, the Vector Space Model (VSM) of Information Retrieval (IR) was suggested, which took its present form in subsequent work (e.g., van Rijsbergen, 1979; Salton and McGill, 1983). Every term is conceived as a separate dimension (axis) of a vector space, in which every document and query can be seen as a — formal rather than physical — vector (of weights). Retrieval is based on different similarity measures used as retrieval functions derived from the dot product (as a measure of count of terms in common) between the query and document vectors.

The view that queries and documents belong to the same and linear space is advantageous from a practical point of view, and all implemented VSM retrieval systems adopt it, although it is not at all clear whether the documents and queries should or not belong to a common space, or the terms (axes) are or not independent from (perpendicular to) each other. This mathematical standpoint (same vector space) also raises a number of basic problems as regards concepts of fundamental importance (similarity, term independence, preference, linearity) in IR (Bollmann-Sdorra and Raghavan, 1993) and of a mathematical nature (Is the vector space used finite or infinite? Do the documents form a linear space theoretically or practically; for example, is the sum of two documents an existing document?) due to which the conceptual basis (using vectors as members of vector spaces) of the VSM of IR has become questionable. In the present paper, a different mathematical formulation of the VSM is proposed. This can serve as a new conceptual basis in which the problems and paradoxes above are dissolved, new paradoxes are not generated (hopefully), and, at the same time, the existing computational practice of the VSM is not invalidated.

2 VECTOR SPACE MODEL OF INFORMATION RETRIEVAL: THE CLASSICAL VIEW

In order to fix the ideas, the mathematical concepts on which the VSM of IR is formally based are reviewed briefly: the notions of linear (or vector) and Euclidean space. Then, the traditional view of the VSM itself, as a formal construct, will be formulated in a concise manner. This will be followed by a short review of the paradoxes and problems to which such a view yields, they justify the quest for a paradox-free formal construct as another formal formulation of the VSM.

2.1 Linear Space

A real linear space (or real vector space) in a set L over the set **Â** of real numbers is a 4-tuple (L, ⊕, ⊗, **Â**), denoted briefly by L, if the following properties hold:

$$a \oplus b \in \mathrm{L}, a \oplus b = b \oplus a, \forall a, b \in \mathrm{L}$$

$$\exists e \in \mathrm{L} \text{ so that } a \oplus e = a, \forall a \in \mathrm{L} \qquad (1)$$

$$\forall a \in \mathsf{L}\ \exists a' \in \mathsf{L} \text{ so that } a \oplus a' = e$$
$$a \oplus (b \oplus c) = (a \oplus b) \oplus c, \forall a, b, c \in \mathsf{L} \qquad (2)$$

$r \otimes a \in \mathsf{L}, \forall r \in \hat{\mathbf{A}}, \forall a \in \mathsf{L}$; and for $\forall r, p \in \hat{\mathbf{A}}, \forall a, b \in \mathsf{L}$ we have:
$$(r + p) \otimes a = r \otimes a \oplus p \otimes a$$
$$r \otimes (a \oplus b) = r \otimes a \oplus r \otimes b \qquad (3)$$

$$(r \times p) \otimes a = r \otimes (p \otimes a)$$
$$1 \times a = a \qquad (4)$$

The elements of a linear space L are traditionally called vectors, and are usually denoted by small bold letters, e.g., $\mathbf{v} \in \mathsf{L}$, while the elements of $\hat{\mathbf{A}}$ are called scalars (in this context). Given the vectors $\mathbf{v}_1, \ldots, \mathbf{v}_m$. The expression $r_1 \otimes \mathbf{v}_1 \oplus \ldots \oplus r_m \otimes \mathbf{v}_m$ is called a linear combination of these vectors, $r_1, \ldots r_m \in \hat{\mathbf{A}}$. If the linear combination is equal to e if and only if $r_1 = \ldots = r_m = 0$, then the vectors $\mathbf{v}_1, \ldots, \mathbf{v}_m$ are said to be linearly independent; they are linearly dependent otherwise. A set of linearly independent vectors forms an algebraic basis (shortly: basis) of L if any vector of the space can be written as a linear combination of them. Every linear space has at least one basis. Each basis contains the same number of vectors, this number is referred to as the dimension of the space. If $\mathbf{b}_1, \ldots, \mathbf{b}_n \in \mathsf{L}_n$ denote basis vectors of an n-dimensional linear space L_n, then every vector $\mathbf{v} \in \mathsf{L}_n$ can be written as a linear combination of the basis vectors: $\mathbf{v} = p_1 \otimes \mathbf{b}_1 \oplus \ldots \oplus p_n \otimes \mathbf{b}_n$, where the scalars $p_1, \ldots, p_n \in \hat{\mathbf{A}}$ are called the coordinates of vector $\mathbf{v}$, notations: v = $(p_1, \ldots, p_n) = [p_1, \ldots, p_n]^T$.

EXAMPLE 1. (i) The n-tuples $(v_1, \ldots, v_n) \in \hat{\mathbf{A}}^n$ form a linear space $(\hat{\mathbf{A}}^n, +, \times, \hat{\mathbf{A}})$ over the set of real numbers with respect to addition: $\mathbf{v} = (v_1, \ldots, v_n)$, $\mathbf{w} = (w_1, \ldots, w_n)$, $\mathbf{v} + \mathbf{w} = (v_1 + w_1, \ldots, v_n + w_n)$, and multiplication: $r \times \mathbf{v} = (r \times v_1, \ldots, r \times v_n)$. (ii) The set of convergent sequences $\langle a_n \rangle_{n \in \hat{\mathbf{A}}} = a_1, a_2, \ldots, a_n, \ldots$ of real numbers form a linear space $(\mathsf{S}, +, \times, \hat{\mathbf{A}})$ with the operations: $\langle a_n \rangle_{n \in \hat{\mathbf{A}}} + \langle a_n \rangle_{n \in \hat{\mathbf{A}}} = \langle a_n + a_n \rangle_{n \in \hat{\mathbf{A}}}$, $r \times \langle a_n \rangle_{n \in \hat{\mathbf{A}}} = \langle r \times a_n \rangle_{n \in \hat{\mathbf{A}}}$; where $\aleph = \{1, 2, \ldots\}$ denotes the set of natural numbers. ♦

2.2 Euclidean Space

Given a real linear space $(\mathsf{L}, \oplus, \otimes, \hat{\mathbf{A}})$. A mapping $\pi: \mathsf{L} \times \mathsf{L} \rightarrow \hat{\mathbf{A}}$ satisfying the properties:
$$\pi(\mathbf{x}_1 \oplus \mathbf{x}_2, \mathbf{y}) = \pi(\mathbf{x}_1, \mathbf{y}) \oplus \pi(\mathbf{x}_2, \mathbf{y});$$
$$\pi(r \otimes \mathbf{x}, \mathbf{y}) = r \otimes \pi(\mathbf{x}, \mathbf{y}) \qquad (5)$$

$$\pi(\mathbf{x}, \mathbf{y}) = \pi(\mathbf{y}, \mathbf{x});$$
$$\pi(\mathbf{x}, \mathbf{x}) > 0 \text{ if } \mathbf{x} \neq e \qquad (6)$$

$\forall\ \mathbf{x}_1, \mathbf{x}_2, \mathbf{y}, \mathbf{x} \in \mathsf{L}, \forall r \in \hat{\mathbf{A}}$, is called a scalar (or inner or dot) product. Instead of $\pi(\mathbf{x}, \mathbf{y})$ the following shorter notations are used: $(\mathbf{x}, \mathbf{y})$, $\mathbf{x} \cdot \mathbf{y}$ or $\mathbf{xy}$.

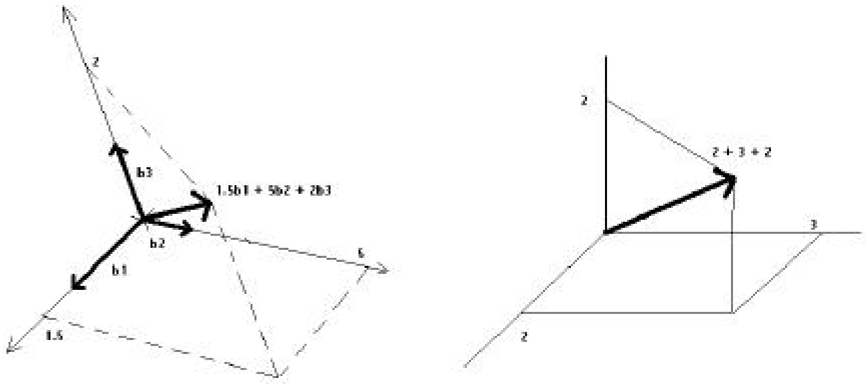
EXAMPLE 2. In the linear space of EXAMPLE 1(i), the following is an inner product: $(\mathbf{v}, \mathbf{w}) = (v_1, \ldots, v_n)\cdot(w_1, \ldots, w_n) = v_1w_1 + \ldots + v_nw_n$. ♦

The length of a vector $\mathbf{v} \in \mathsf{L}$ is denoted by $||\mathbf{v}||$, and defined as $||\mathbf{v}|| =_{\text{def}} (\mathbf{v}, \mathbf{v})^{0.5}$. A vector $\mathbf{v}$ has unit length if $||\mathbf{v}|| = 1$. A distance δ between two vectors $\mathbf{v}$ and $\mathbf{w}$ can be defined as $\delta(\mathbf{v}, \mathbf{w}) =_{\text{def}} ||\mathbf{v} \oplus (-1)\otimes\mathbf{w}||$. The measure of the angle between two vectors $\mathbf{v}, \mathbf{w} \neq \mathbf{0} \in \mathsf{L}$ is a real number φ satisfying the property: $\cos \varphi = \mathbf{v}\cdot\mathbf{w} / (||\mathbf{v}||\times||\mathbf{w}||)$. Two vectors $\mathbf{v}$ and $\mathbf{w}$ are orthogonal if $\cos \varphi = 0$.

A real linear space $(\mathsf{L}, \oplus, \otimes, \mathbf{\hat{A}})$ with inner product is called an (real) Euclidean space, notation: E. Any n-dimensional Euclidean space E_n has an orthonormal (i.e., orthogonal and unit lengths basis vectors) basis (there may be other bases, too, which need not be orthogonal or with unit lengths basis vectors). A common orthonormal basis is: $\mathbf{e}_1 = (1, 0, \ldots, 0)$, $\mathbf{e}_2 = (0, 1, 0, \ldots, 0)$, …, $\mathbf{e}_n = (0, 0, \ldots, 1)$.

EXAMPLE 3. The linear space of EXAMPLE 1(i) with the inner product of EXAMPLE 2 is an n-dimensional Euclidean space. The Euclidean norm of a vector $\mathbf{v} = (v_1, \ldots, v_n)$ is $||\mathbf{v}|| = \sqrt{\sum_{i=1}^{n} v_i^2}$, while the Euclidean distance of two vectors $\mathbf{v} = (v_1, \ldots, v_n)$ and $\mathbf{w} = (w_1, \ldots, w_n)$ is $||\mathbf{v} + (-1) \times \mathbf{w}|| = \sqrt{\sum_{i=1}^{n} (v_i - w_i)^2}$. ♦

Figures 1 visualises the concept of Euclidean space.



(a) (b)

Figure 1. (a) Visualisations of E_3. The vectors **b**1, **b**2 and **b**3 form a basis. The basis vectors do not have unit lengths. They are represented on three coordinate axes, which are not perpendicular. A vector **v** given by the linear combination $1.5\mathbf{b}_1+5\mathbf{b}_2+2\mathbf{b}_3$is represented. (b) Visualisations of E_3 with orthonormal basis. A vector given by the linear combination $2\mathbf{e}_1+3\mathbf{e}_2+3\mathbf{e}_3$ is represented.

2.3 Vector Space Model of Information Retrieval

Given a set $T = \{t_i \mid i = 1,\ldots,n \in$ **À**$\}$ of elements (criteria) called *terms* (all terms are different, so they may form a set), and entities D_j, $j = 1,\ldots,m \in$ **À** called *documents* (there may be identical documents, so they may not always form a set); **À** denotes the set of natural numbers. Every document is associated — conceived as or characterised by — a finite sequence of terms, i.e., $D_j \sim \langle t_k\rangle_j = t_{k_1},\ldots,t_{k_j}$, which, in set notation, is $\{(t_k, f_k) \mid t_k$ belongs to $\langle t_k\rangle_j$, f_k denotes the multiplicity of occurrence of $t_k\}$.

EXAMPLE 4. $D \sim \langle$sun, sun, sky, sky, sky$\rangle \sim \{$(sun, 2), (sky, 3)$\}$ ♦

Let $w_{ij} \in$ **Â** denote the *weight* of term t_i for document D_j, i.e., a numerical measure of the degree to which a term pertains to or characterises a document. Thus, a matrix $W = (w_{ij})_{n\times m}$, called the *term-by-document matrix*, is obtained. A number of manual as well as automatic methods have been proposed to compute the weights. The automatic ones are usually expressed using several separate formulas (Belew, 2000; Berry and Browne, 1999). Those generally accepted are — formally — combined here into one concise formula as follows:

$$w_{ij} = \frac{w'_{ij}}{(\max\limits_{1\le k\le n} w'_{kj})^{n_1} \cdot \sqrt[n_3]{\sum\limits_{k=1}^{n} \left(w'_{kj}\right)^{n_2}}}, \qquad \text{where} \tag{7}$$

$$w'_{ij} = \left(\ln \left(f_{ij}^{\,l_1} \cdot e^{\frac{l_2-1}{2} + \frac{f_{ij}}{l_3 \cdot \max\limits_{1\le k\le n} f_{kj}}} \right)\right) \cdot \left(\frac{g_1}{F_i} + \left(\log \frac{m - g_3 F_i}{F_i} \right)^{g_2} \right) \tag{8}$$

where f_{ij} denotes the number of occurrences of term t_i in $\langle t_k\rangle_j$, and F_i the number of sequences $\langle t_k\rangle_j$ in which the term t_i occurs. Every document D_j is represented by a vector $\mathbf{w}_j = (w_{1j}, \ldots, w_{nj}) \in E_n$ of weights. $\lambda_1, \lambda_2, \lambda_3, \gamma_1, \gamma_2, \gamma_3$ ν_1, ν_2, ν_3 are real parameters. Table 1 shows significant values of these parameters yielding usual formulas developed thus far.

Table 1. Weghting schemes matrices.

l_1	l_2	l_3	Scheme	g_1	g_2	g_3	Scheme	n_1	n_2	n_3	Scheme
0	1	1	MaxNorm	0	1	0	IDF	0	1	1	Sum
0	2	2	Augmented	0	2	0	Square	0	2	2	Cos
1	3	∞	Log	0	1	1	Prob	0	4	1	4th
				1	0	γ_3	Freq	1	0	∞	Max

Of course, w_{ij} can be, as a special case, just f_{ij} (binary or not). We should note that not all possible values and combinations for these parameters are used in practice, and certain combinations of the parameters values are preferred; for example: w = 'Augmented' for technical literature, w = 'MaxNorm' × 'IDF' for encyclopaedias.

Let Q denote a *query*, and $\mathbf{q} = (q_1, \ldots, q_n)$ the corresponding query vector (whose elements can be calculated using Form. 8 or some other method). Both vectors $\mathbf{w}_j$ and $\mathbf{q}$ belong to the same n-dimensional Euclidean space E_n. The *relevance* of a document D_j relative to Q is given by a real valued *retrieval function* $\rho(\mathbf{w}_j, \mathbf{q})$ based on similarity measures. Several such functions have been developed, and they are usually expressed using several separate formulas. Those generally accepted (Meadow et al., 1999) are combined here into one concise formula as follows:

$$\rho = \frac{\mathbf{w}_j\mathbf{q}}{(\|\mathbf{w}_j\|\cdot\|\mathbf{q}\|)^{1/a}\cdot(2^{c-b}(|\mathbf{w}_j| + |\mathbf{q}|) - \mathrm{c}\cdot\mathbf{w}_j\mathbf{q})^b} \qquad (9)$$

where $|\mathbf{x}|$ denotes the sum of the elements of vector $\mathbf{x}$. a, b and c are real parameters. Table 2 shows significant values of these parameters yielding usual formulas developed thus far.

Table 2. Retrieval function matrix.

a	*b*	*c*	**Name of Retrieval Function**
∞	0	*c*	Dot product
2	0	*c*	Cosine measure
∞	1	0	Dice' coefficient
∞	1	1	Jaccard's coefficient

There are other particular weighting schemes and retrieval functions, too. Mathematically undefined cases, such as, e.g., 0^0, or all f_{ij} are null, etc., should obviously be excluded on the ground that they can, in practice, be detected and excluded before any computation should take place.

To summarise, the following formal and concise definition can be given for the VSM. Figure 2 is a possible visualisation of the VSM.

*D*EFINITION 1. The VSM of IR is the 5-tuple VSM = (T, D, W, Q, ρ). ♦

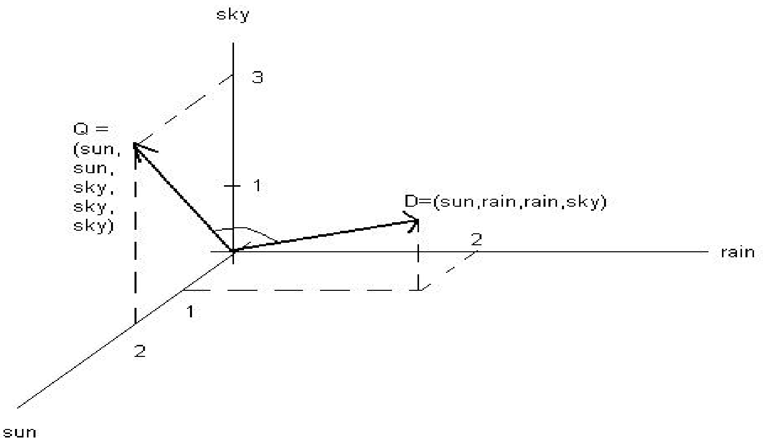


Figure 2. The set T of terms is T = {sun, rain, sky}. A document D ~ <sun, rain, rain, sky> and a query Q ~ <sun, sun, sky, sky, sky> are shown. The corresponding vectors of weights are (using a frequency weighting scheme): $\mathbf{w} = (1, 2, 1)$, $\mathbf{q} = (2, 0, 3)$. The angle between these two vectors is a measure of similarity between the document and query.

2.4 Paradoxes of the Vector Space Model of Information Retrieval

In the VSM, both the document and query vectors belong to the same n-dimensional linear (Euclidean) space. This mathematical view implies that properties like linearity, independence, or preference should simultaneously hold for both. Also, it is natural to assume that

(i) similar queries give comparable performance,
(ii) a query that is more similar to the perfect query yields increased retrieval performance,
(iii) queries corresponding to relevant documents lead to better performance than those corresponding to nonrelevant documents.

However, this is not the case. It is well known and widely accepted that the VSM, in general, has problems from the very start; for example, the assumption that the index terms are/should be independent from each other is not always a realistic one, nor is it clear whether the query and document are/should/may be or not entities of the same type (and thus belonging or not to a common space). The paradoxes that can be shown mathematically in the classical view of the VSM are as follows.

1. A query that is closer to the perfect query does not necessarily yield better retrieval performance.
2. Inserting terms into a document does not decrease its similarity to queries, whereas the insertion of a term into a query may yield decreased retrieval performance.
3. The similarity of documents and queries does not necessarily reflect the same kind of preference on both documents and queries.
4. Term independence does not necessarily hold simultaneously for documents and queries (it is possible to have term independence in documents but not in queries, and vice versa).
5. For a finite document space, which is the case in practice, a linear structure on queries implies that all queries are equally good. In other words, there is not (except the former trivial case) a linear structure on the space of queries with weighted terms (i.e., potentially there are infinitely many queries).
6. Weighted retrieval is incompatible with term independence in the query space.

In addition to the paradoxes above, further problems arise. For example, given that the n-dimensional Euclidean space is a Banach space, too, i.e, a linear space with a norm in which every Cauchy sequence is convergent and vice versa, the following questions arise:

(a) Is there a sequence $\langle \mathbf{q}_n \rangle_{n \in \grave{\mathbf{A}}}$ of query vectors in the space of potential queries that is a Cauchy sequence?
(b) How can such a sequence be constructed?
(c) If such a sequence exists, then it converges to a query vector $\mathbf{q}$, i.e., $\lim \mathbf{q}_n \to \mathbf{q}$. One would expect then that the corresponding sequence

$\langle \mathbf{w}_n \rangle_{n \in \text{À}}$ of retrieved documents converge to the document vector $\mathbf{w}$ which corresponds to $\mathbf{q}$, i.e., lim $\mathbf{w}_n \to \mathbf{w}$. This is, however, questioned by point 1 above in the case when $\mathbf{q}$ corresponds to the optimal query ($\mathbf{w}$ may not be the most relevant document).

Due to the paradoxes and problems above, the underlying idea of the VSM according to which both the documents and queries are elements of the same linear space (and thus retrieval is based on similarity), is questionable: it yields meaningless cases in IR. In other words, the usual classical requirements (same space, Euclidean space, term independence) are too strong, they are not needed actually. One way to avoid the paradoxes that such assumptions generate is to view the queries and documents as belonging to two separate spaces (Bollmann-Sdorra and Raghavan, 1998); these spaces may or may not be linear spaces, they should, however, bear an ordered structure (in order to model preference or relevance relations). In this view, the paradoxes seem to be cancelled under the realistic assumption that the document space is finite; also, the structure of the query space can be different from that of the document space have (which reflects the intuition according to which the conceptual 'background' of a person who is asking can indeed be very different from the person's who is answering). Such situations trigger the challenge to find another mathematical formulation of the VSM that aims at being paradox-free and at being formally simpler (not implying a complex algebraic or geometric structure and its study of queries and documents). On the other hand, however, because other models of IR do not or hardly lead to considerably better retrieval performances than those obtained in implemented practical VSMs, such a newer mathematical foundation of the VSM should

- not assume the same and linear space for documents and queries (to remove the paradoxes so generated),
- should not invalidate the computational practice of the VSM (formulas used in practice are kept).

3 PROBABILITY SPACE AND AMOUNT OF INFORMATION

In this part, the concept of a probability space is briefly recalled (Kolmogoroff, 1933). This will be followed by a review of the concept of Shannon (1948) information and its measure. These notions will then be used in Part 4 to re-formulate the VSM as a special probability space.

3.1 Probability Space

Given a set Ω called universe, and let its elements be called elementary events. A set $\Im \subseteq \wp(\Omega)$ is called a σ-algebra if $\Omega \in \Im$, and $\forall A, B \in \Im$ we have $A \cap B \in \Im, A \cup B \in \Im, \Omega \setminus A \in \Im$. A probability measure is a function $P: \Im \to [0; 1]$ satisfying the properties $P(\Omega) = 1$, $A \cap B = \emptyset \Rightarrow P(A \cup B) = P(A) + P(B)$, $\forall A, B \in \Im$. The triple $(\Omega, \Im, P) = \Psi$ is called a probability space.

3.2 Amount of Information

Given alternatives (events) E_j, and let p_j denote the probability to select alternative (probability of occurrence of event) E_j, $j = 1,\ldots,m \in$ **À**. Information is conceived as different from meaning, and defined as one's freedom of choice. A measure H for information is

$$H = -k\sum_{j=1}^{m} p_j \log_2 p_j \tag{10}$$

where k is a positive constant, which amounts to a choice of a unit of measure. The quantity H satisfies the following properties:

a) The amount of information is zero if and only if exactly one alternative is selected:

$$\lim_{\substack{p_k \to 1 \\ p_j \to 0, \forall j \neq k}} H = 0 \qquad \Leftrightarrow \qquad p_k = 1; \quad p_j = 0, \ \forall j \neq k \tag{11}$$

b) The amount of information is maximal and equal to $\log_2 m$ if all p_j are equal to $1/m$:

$$P_j = \frac{1}{m}, \quad j = 1,\ldots,m \qquad \Rightarrow \qquad H = \max H = \log_2 m \tag{12}$$

c) The closer all p_j to each other the larger the amount of information:

$$P > P' \qquad \Rightarrow \qquad H < H';$$

$$\text{where } P = \sum_{j=1}^{m}\left(\frac{1}{m} - p_j\right)^2 \tag{13}$$

Thus, for example, if one has total freedom to choose from two alternatives then the amount of information associated to this situation is considered to be unity (1 bit). Although it is not stated explicitly, from properties (11) and (12) it follows that $\sum_{j=1}^{m} p_j = 1$; otherwise H could, for example, be null for $p_j = 1$, $j = 1, \ldots, m$, too (which is mathematically true), and the — mathematically correct — maximum of H would be reached for $p_j = 2^{-\frac{1}{\ln 2}} = 0.368$, $j = 1,\ldots, m$ (all partial derivates are null).

The property (13) will play an important role in the present context, thus, without restricting its general validity, let us have a closer look at the case $m = 2$.

THEOREM 1. $P < P' \Rightarrow H > H'$.

Proof. Let $p_1 = p$, $p_2 = q$. The condition $P < P'$ means that $(0.5 - p)^2 + (0.5 - q)^2 < (0.5 - p')^2 + (0.5 - q')^2$. We can assume that $p' = p + a$, $q' = q - a$. From this we obtain $p - q + a > 0$. From $q - a < p$ and $\log(1/(q - a)) > \log(1/p)$ it follows that $p \cdot \log(1/p) > (q - a) \cdot \log(1/(q - a))$; from $q < p + a$ and $\log(1/q) > \log(1/(p + a))$ it follows that $q \cdot \log(1/q) > (p + a) \cdot \log(1/(p + a))$. Hence $p \cdot \log(1/p) + q \cdot \log(1/q) > p' \cdot \log(1/p') + q' \cdot \log(1/q')$, i.e., $H > H'$. ♦

In other words, if the probabilities p_j change such that they deviate more from $1/m$ (some are closer to 1 whilst the others closer to null) the amount of information (or uncertainty) becomes smaller, i.e., the freedom to select becomes more restricted. This entitles us to introduce the following

*D*EFINITION 2. An operation (procedure, process, mechanism) which spreads the probabilities p_j from $1/m$ is called an *information decreasing operation* (IDO). ♦

Thus, any operation which constrains the freedom (and thus reduces the uncertainty) to select is an IDO.

EXAMPLE 5. (i) Numerical minimisation of a function based on the gradient method: the freedom to select a direction to follow is decreased because not any direction may be followed, only that given by the gradient. (ii) Breadth-first search algorithm: the freedom to move to a next vertex is constrained because a downward walk is not allowed as long as there are unexplored breadth vertices (from the current vertex, we may not jump to any other vertex we want). (iii) A person on diet does not (or should not) have total liberty to choose the bread or meat he/she would like. ♦

In cases like these, the freedom of choice to select from alternatives is constrained, some of the alternatives are/should be selected with higher probabilities in the detriment of the others, and thus the total amount of information (and uncertainty) associated to the selection situation as a whole is decreased.

4 VECTOR SPACE MODEL OF INFORMATION RETRIEVAL: INFORMATION DECREASING PROBABILITY SPACE

In this part, based on the concepts of a probability space and IDO, the notion of the VSM will be given a different (from the traditional one) mathematical formulation: it will be shown that the VSM may be conceived as a probability space that decreases the amount of information, i.e., the VSM is an IDO probability space. This view is independent of whether or not the documents and queries are vectors or belong to the same or different space. At the same time, the computational practice (formulas used for computations) of the VSM is not invalidated.

We first prove that a probability space having its probability measure defined in a certain way is an IDO.

*L*EMMA. Let $\Psi = (\Omega, \Im, P)$, $|\Omega| = m$, denote a probability space with the probability measure P defined as follows:

$$P(X) = \begin{cases} \dfrac{r_j}{\sum_{k=1}^{m} r_k}, X = X_j \in \Omega \\ \\ P'(X), otherwise \end{cases}$$

where not all ρ$_j$ are equal to each other, i.e., $\exists\ k \neq s$ such that $\rho_k \neq \rho_s$. (An explicit formula for *P'* does not play any role in this context.) Then the probability space Ψ is and IDO.

Proof. The space Ψ is an IDO if it spreads the probabilities from 1/*m* (Def. 2), i.e., we have to show that

$$P = \sum_{j=1}^{m}\left(\frac{1}{m} - p_j\right)^2 > 0, \qquad \text{where } p_j = \frac{r_j}{\sum_{k=1}^{m} r_k}.$$

We can write

$$P = \sum_{j=1}^{m}\left(\frac{1}{m} - \frac{r_j}{\sum_{k=1}^{m} r_k}\right)^2 = \frac{\sum_{j=1}^{m} r_j^2}{\left(\sum_{k=1}^{m} r_k\right)^2} - \frac{1}{m} > 0 \qquad \Leftrightarrow$$

$$\Leftrightarrow\ m\sum_{j=1}^{m} r_j^2 > \left(\sum_{k=1}^{m} r_k\right)^2$$

Because $\exists\ k \neq s$ such that $\rho_k \neq \rho_s$ we have

$$m\sum_{j=1}^{m} r_j^2 - \left(\sum_{k=1}^{m} r_k\right)^2 = \sum_{\substack{k=1\\ s=k+1}}^{\substack{m\\ m-1}} (r_k - r_s)^2 > 0 \qquad \blacklozenge$$

As a paradox-free formulation of the VSM, it can now be shown that it may be conceived as a different mathematical construct (i.e., not based on a Euclidean space containing both documents and queries), namely a probability space that decreases the amount of information, as follows:

*T*HEOREM 2. The VSM of IR is an IDO probability space Ψ.

Proof. Given hence (Def. 1) a VSM = (*T*, *D*, *W*, *Q*, ρ). The query *Q* is also associated a sequence ⟨**q**⟩ = $q_1, \ldots, q_n$ of weights. The matrix *W* is viewed as a matrix (sequences simply arranged this way), its columns are not conceived as vectors of any linear space (even if they were this is irrelevant here now), and the same holds for the query *Q*, too. Whether the weights are computed automatically or established in some other way is irrelevant. The measure of relevance given by the value of the retrieval function ρ(**w**$_j$, **q**) = ρ$_j$ represents a degree of the choice of document D_j as a response to query *Q*. The value ρ$_j$ can always be calculated, independently of whether its parameters are or not viewed as vectors. The higher the value of ρ$_j$ the higher the chance of document D_j to be selected as an answer. Thus, a sequence ⟨**r**⟩ = $\rho_j, \ldots, \rho_m$ is defined (this is viewed just as a sequence and not as a vector; even if it was it would be irrelevant here now), which represents the choices of all documents relative to query *Q*. (The case when all the ρ$_j$ are null can be excluded as trivial.) Using the sequence ⟨**r**⟩, the following sequence ⟨***P***⟩ = $p_1, \ldots, p_m$, where

$$p_j = r_j \Big/ \sum_{k=1}^{m} r_k,\ j = 1, \ldots, m,$$

is defined, which can be viewed as the probabilities to select the documents as answers in the following probability space

$$\Psi = (\Omega, \Im, P), \quad \Omega = \{D_1, \ldots, D_m\}, \qquad \text{where}$$

$P(D_j) = p_j$, and $P(X) = P'(X)$ if $X \neq D_j$.
Hence (Lemma), the amount

$$H = -\sum_{j=1}^{m} p_j \log_2 p_j$$

of the Shannon information corresponding to the retrieval situation as a whole is decreased, and thus the VSM is and IDO. ♦

The following example illustrates the view according to which the VSM is an IDO probability space.

EXAMPLE 6. Let us consider three documents: D_1, D_2 and D_3, $m = 3$, and two terms: t_1 and t_2, $n = 2$. Let the frequencies of terms in documents be as follows (i = 1, 2; j = 1, 2, 3): $W = \begin{bmatrix} 2 & 1 & 1 \\ 0 & 3 & 2 \end{bmatrix}$. Let Q denote a query, and the corresponding term frequencies be (0, 1). For computational convenience, we will use matrix notation.

(a) If the retrieval function ρ is the dot product, then the chances ρ_1, ρ_2 and ρ_3 are $W^T\mathbf{q}$ = $\begin{bmatrix} 2 & 0 \\ 1 & 3 \\ 1 & 2 \end{bmatrix} \times \begin{bmatrix} 0 \\ 1 \end{bmatrix} = \begin{bmatrix} 0 \\ 3 \\ 2 \end{bmatrix}$, whilst the corresponding probabilities are $\begin{bmatrix} 0 \\ 0.6 \\ 0.4 \end{bmatrix}$. The associated amount of information is decreased to 0.971 from the maximum $\log_2 3 = 1.585$.

(b) If the retrieval function is the Cosine measure and the weights are max normalised, then the chances ρ_1, ρ_2 and ρ_3 are $\begin{bmatrix} 0 \\ 0.923 \\ 0.847 \end{bmatrix}$, whilst the corresponding probabilities are $\begin{bmatrix} 0 \\ 0.521 \\ 0.479 \end{bmatrix}$. The associated amount of information is decreased to 0.999 from the maximum $\log_2 3 = 1.585$. ♦

Based on Theorem 2, the ideal (when, ideally, the similarity measure does indeed express relevance) VSM of IR may be formulated as an information optimisation problem as follows:

DEFINITION 3. The ideal VSM of IR aims at finding weights and retrieval function such that the amount of information be a local minimum given terms, documents and queries: $\min_{W,r} H$, given T, D, Q. ♦

5 EXPERIMENTS

Experiments were carried out with three text collections to monitor the variations of the amount of information. Two of these text collections were the standard test collections MEDLINE and TIME, where index terms were obtained automatically using standard techniques (stop list, Porter-stemming). Because these test collections are widely known, they are not described and details are not presented (their statistics is as usual).

The third collection of texts, referred to as BELIEFS, contained 2704 Hungarian belief texts. The choice of this collection was motivated by two reasons: it was not a standard one, and it was very special. Different spellings were used: contemporary Hungarian (for instance, "Ha kis gyermeknek komoly baja van, akkor szenes vizzel mossák meg. A meleg vizbe 9 drb. szenet tesznek, megkenik a vizzel a gyermek homlokát és ezt mondják: Ha férfi, kalap alá; ha leány párta alá; ha asszony fejkötô alá, az atya, fiú, szentlélek nevében. Amen."), mixture of older Hungarian spelling and dialect (for example, "Ha a tehenet merrontya a boszorkány, vësznek egy új fëlliteres cserepbëgrét; abba belëtësznek ecs csomaócskát a tehen gannajjábó. Azután szöget vernek a kény belsejébe s erre felakasztyák a bëgrét. Etteô aszt meggyön a tehen haszna."). Also, many different word forms were used. Due to these characteristics the text pre-processing operations had been carried out manually. A number of 1,551 stop words (e.g., pronouns, adverbs, articles, attributes, verbs, present participles, rarely used chemical words, as well as conjugated/declined forms) were identified manually as baring no or very little significance for beliefs, and gathered in a list (for example, different declined forms of the personal pronoun "aki" meaning "who": "aki, akié, akiébe, akiért, akihez, akijé, akik, akiknek, akin, akinek, akinél, akire, akirôl, akit, akitôl, akivel, akki, akkinek, akkire"). After the automatic removal of the stop words there remained 14,286 word forms. The word forms were then stemmed manually. For example, the declined word forms: "csont, csontjával, csontja, csontig, csontok, csontjait, csontnak, csontokat, csontra, csontjai, csontom, csont, csonton, csontját, csontokbúl, csontot, csonttal" were all stemmed to "csont" meaning "bone". A further difficulty stemmed from the very many composed words, which are typical for the Hungarian language (just like in German, for instance). A special difficulty was posed by old homonym words which are not used anymore; for example, the words "betyöleges, bíszbányosok" were replaced by the word "varázs" meaning "magic". The result was a number of 2,607 terms in a correct contemporary spelling, which were used as index terms for the belief texts. The average number of terms per text was 15. Table 3 shows the results of test runs. Three weighting schemes were used for each text collection: frequency, MaxNorm, and CosNorm. Four measures were calculated in each case: Cosine, Dot product, Dice's coefficient, and Jaccard's coefficient. The amount of information *H* was calculated in each case. Also, the average values of *H* for each weighting scheme and similarity measure was computed. The percentages of the decrement of *H* is shown in every case, too.

Table 3. Decrements of the amount of information in VSM as an IDO probability space.

DATA BASE	**Weighting scheme**	**Amount of information *H* if retrieval measure is:**				**Average values of *H***	H_{max}	***H* decreased by (%)**
		Cosine	*Dice coeff.*	*Jaccard coeff.*	*Dot prod.*			
MEDLINE	Frequency	9.943	9.599	9.330	9.330	9.551	10.013	4.6
	MaxNorm	9.943	8.644	7.520	7.520	8.41		16
	Cos norm.	9.943	9.969	9.969	9.943	9.956		0.57
	Average values of *H*	9.943	9.404	8.94	8.931			
	***H* decreased by (%)**	1	6.1	10.8	10.9			
TIME	Frequency	8.632	8.151	7.808	7.809	8.1	8.725	7.16
	MaxNorm	8.632	6.184	4.726	4.727	6.067		30.46
	Cos norm.	8.632	8.675	8.675	8.632	8.653		0.82
	Average values of *H*	8.632	7.67	7.07	7.06			
	***H* decreased by (%)**	1.1	12.1	19	19			
BELIEFS	Frequency	11.321	10.959	10.793	10.813	10.971	11.401	3.77
	MaxNorm	11.321	10.790	10.275	10.275	10.665		6.45
	Cos norm.	11.321	11.340	11.34	11.321	11.33		0.62
	Average values of *H*	11.321	11.03	10.8	10.8			
	***H* decreased by (%)**	0.8	3.3	5.3	5.3			

It can be seen that the amount of uncertainty was decreased to the greatest extent using the MaxNorm weighting scheme, and less using the CosNorm scheme. As regards the similarity measures, the Dot product as well as the two coefficients (Dice’s and Jaccard’s) performed — perhaps unexpectedly — more reduction compared to the Cosine measure. In a future research, a possible correlation between recall and precision, and the decrement of *H* is planned to be analysed.

6 DISCUSSION AND CONCLUSIONS

In this paper, after reviewing the concepts of linear and Euclidean spaces, the traditional view of the VSM was recalled in a compact form. The view according to which both the documents and queries are viewed as vectors belonging to the same Euclidean space raised paradoxes and problems which questioned its validity, although the practice and relevance effectiveness of the VSM could hardly be outperformed. Thus, in order to remove the paradoxes, and to keep the computational practice, in the present paper it was suggested to conceive the VSM of IR as an IDO, i.e., a probability space that decreases the amount of Shannon-information. This view may serve as a paradox-free formal foundation of the VSM. At the same time, it was possible to formulate the concept of an ideal VSM as an information optimisation problem.

In the traditional view of the VSM, both the document and query weights are conceived as vectors of the same Euclidean space E_n. Based on the Dot product or Cosine measure, the nice idea to view the VSM as a tensor of rank 2 (Lanczos, 1970; Simmonds, 1982) may arise (the inner product and norm are elementary tensors). This is very appealing, but the paradoxes (Part 2.3) do not allow it. One might then think of the query vector as belonging to the space E_n, and the term-by-document matrix as the components of a tensor of rank 2, to avoid the usage of the same space for queries and documents. But then the effect of the tensor (i.e., the probabilities for the documents to be answers) should belong to another linear space L_n having the same dimension n. This view, too, is again excluded by two factors: (i) paradox 5 (Part 2.3), and (ii) the dimension of the space L cannot be, in general, n: the number of documents is not equal, in general, to that of terms.

At a first look, it may sound strange to say that the amount of *information is decreased* in the IDO-based view of the VSM of IR, when we know that the aim of IR is to *find information*. The concept of information used in an IDO is not identical to that used in IR, where the word information is used in a more complex way. The concept of Shannon information is equivalent to uncertainty, whereas retrieving information in the classical interpretation of the VSM means finding a document whose content (reflected in its terms) is close enough to the content of the query (expressed by its terms). When the VSM is viewed as an IDO probability space, then the amount of uncertainty to select documents is reduced, and thus a need for information is being reduced. (In other models of IR, retrieval is defined in different ways, and there is not consensus on whether information or content or meaning or something else is being or should be retrieved in IR.)

In the view suggested in the present paper (VSM as IDO probability space), the weights associated to documents and queries are, for expression convenience, arranged and represented as matrices. Whether the sequences of weights (columns of matrices) are viewed as belonging to any space or not is irrelevant in this context. However, because the same set of terms *T* is used to generate the weights, one may think that, in fact, the same conceptual space is being used (in spite of the fact that the presence of a space is not explicitly stated). Documents and queries should, in fact, be — and are indeed — conceived as different entities, although they both can be characterised by common criteria (just as, for example, fishes, trees, and uranium can be characterised by mass and lifetime, which does not mean that they necessarily belong to the same conceptual space). For computational convenience, matrix representations are used, which can be misleading.

Experiments were performed to observe the decrement of the quantity *H* of information. It was found that this varied between 0.57 % and 30.46 % depending on the weighting schemes, and between 0.8% and 19% depending on the similarity measures used. The fact that the different similarity measures led to different results was not totally unexpected because, as it is well-known, on the one hand they give different similarity values, and, on the other hand, they do not preserve the rank ordering in general. However, it is a bit surprising that the Cosine-based methods yielded the worst performances as regards both weighting schemes and similarity measures. These aspects justify a future research into relationships between recall-precision and quantity *H*.

The emphasis in this paper is that the VSM may be viewed as a probability space in which documents are selected based on probabilities given by similarity measures. The amount of Shannon information is associated only with the intention to show that the retrieval system does operate, and not with the intention to have an effectiveness measure of such a system. Whether a document is retrieved or not depends on ranking and threshold rather than on the effective values of similarities. The different degrees of information reduction in the experiments should not be interpreted as expressing how well or badly the different similarity measures perform. Perhaps mainly theoretically, it may happen that the similarity measure takes on exactly the same values for all the documents; then, depending on the particular value of the threshold, either all or none documents are returned. Such a situation is not typical in real cases, where some documents have higher similarity values and hence higher chances to be returned than others — it is this fact that the amount of associated information (IDO) is meant to reflect. Why different similarities yield different reductions, and whether this has or not any significance may represent a future research topic.

The view suggested in the present paper according to which the VSM is an IDO probability space raises the question of whether this is related to the Probabilistic Model (PM) of IR. In the PM, and in the IDO-VSM (suggested in this paper), too, the choice of a document is based on a probability associated to it (the computational technicalities are different). Thus, in principle, the two views, PM and IDO-VSM, seem to make use of the same mathematical construct (probability), and this idea harmonises with an earlier possibility to have a unified formal definition for the VSM and PM (Dominich, 2001). Also, it would be worth researching into possible links between the reduction of uncertainty introduced in this paper and the concept of uncertainty in the logical models of IR.

Using formula (9), a possible explicit form of the optimisation (ideal!) problem of Def. 3 can be obtained as a nonlinear optimisation problem in which a matrix and a function should be found such that the amount H be (locally) minimal.

The importance of giving a paradox-free formal foundation for the VSM of IR is supported not only by a formal elegancy or the history of this model and its impact on academic research. However strange it may sound, and perhaps contrary to common opinion, but, from a formal point view, the VSM is being used in, e.g., current commercial Web searching, too. For example, Altavista's retrieval may be conceived as a (classical) VSM as follows:

(i) Web pages, which correspond to documents D_j, are seen as segments S_{j1}, S_{j2}, S_{j3} = *Title_first_eight_words*, *Rest_of_Title*, *Body_of_page*;

(ii) Every segment is conceived as a sequence of terms (terms are stored in the inverted file and index score) $S_{ju} \sim \langle t_k \rangle_{ju}$;

(iii) Every segment is assigned a sequence $\langle w_k \rangle_{ju}$ of weights obtained by multiplying the corresponding score s_k of the Index Score (the Index Score is derived based on the Zipf Law, for example "where: 0.66, logarithmic: 275") with a segment specific constant C_{ju} ($C_{j1} = 5$, $C_{j2} = 4$, $C_{j3} = 1.5$), i.e., $C_{ju}s_k$; Thus, the term-by-document matrix is a collection of columns grouped by three corresponding to segment;

(iv) Query terms are assigned unit weights $\langle q_k \rangle$;

(v) Terms not occurring are assigned weights zero;

(vi) The rank value (similarity) of a Web page is obtained by the sum of the Dot products between the row matrices of segments and column matrix of the query, i.e., $\Sigma_u(\mathrm{w}_{ju}, \mathrm{q})$.

This shows that the role and importance of the VSM of IR has not diminished over time (perhaps contrary to current views).

7 Acknowledgements

The Author wishes to acknowledge the support of grants National Research Fund OTKA T037821, Hungary; Ministry of Education OM 359/2001, Hungary; and Hungarian Academy of Sciences AKP 2001-140. Thanks are due to Dr. Sándor Darányi for providing the BELEIFS texts, and to M.Sc.students Tamás Kiezer and Zoltán Szlávik for helping obtain the term-by-document matrices.

References

Belew, K.B. (2000). *Finding Out About.* Cambridge University Press.

Berry, M.W. and Browne, M. (1999). *Understanding Search Engines – Mathematical Modeling and Text Retrieval.* SIAM, Philadelphia.

Bollmann-Sdorra, P. and Raghavan, V. V. (1993). On the Delusiveness of Adopting a Common Space for Modelling IR Objects: Are Queries Documents? *Journal of the American Society for Information Science*, **44**(10): 579-587.

Bollmann-Sdorra, P. and Raghavan, V. V. (1998). On the Necessity of Term Dependence in a Query Space for Weighted Retrieval. *Journal of the American Society for Information Science*, **49**, November: 1161-1168.

Dominich, S. (2001). *Mathematical Foundations of Information Retrieval.* Kluwer Academic Publishers, Dordrecht, Boston, London.

Kolmogoroff, A. (1933). *Grundbegriffe der Wahrscheinlichkeitsrechnung.* Julius Springer, Berlin.

Lanczos, C. (1970). *Space Through The Ages*. Academic Press, London.

Luhn, H. P. (1959). Keyword-in-Context Index for Technical Literature. In: Hays, D. G. (1966, ed.) *Readings in Automatic Language Processing.* American Elsevier Publishing Company Inc., pp. 159-167.

Meadow, C.T., Boyce, B.R. and Kraft, D.H. (1999). *Text Information Retrieval Systems.* Academic Press, San Diego, San Francisco, New York, Boston, London, Sydney, Tokio.

Salton, G. (1965). Automatic Phrase Matching. In: Hays, D.G. (1966, ed.) *Readings in Automatic Language Processing.* American Elsevier Publishing Company, Inc., New York, pp. 169-188.

Salton, G. and McGill, M. (1983). *Introduction to Modern Information Retrieval.* McGraw Hill, New York.

Salton, G., Wong, A. and Yang, C.S. (1975). A Vector Space Model for Automatic Indexing. *Communications of the ACM*, vol. 18, pp. 613-620.

Shannon, C. (1948). *A Mathematical Theory of Communication.* The Bell System Technical Journal, vol. 27, pp. 379-423, 623-656, July, October.

Simmonds, J.G. (1982). *A Brief on Tensor Analysis*. Springer Verlag.

Van Rijsbergen, C.J. (1979). *Information Retrieval.* Butterworth, London.

Zipf, G. (1949). *Human behavior and the principle of least effort*. Addison-Wesley, Cambridge, MA.