An Entropy-Based Interpretation of Retrieval Status Value-Based Retrieval, and Its Application to the Computation of Term and Query Discrimination Value

Sándor Dominich, Júlia Góth, Tamás Kiezer, and Zoltán Szlávik
Department of Computer Science, University of Veszprém, Egyetem u. 10, 8200 Veszprém, Hungary.
E-mail: {dominich, goth, kiezer, szlavik}@dcs.vein.hu

The concepts of Shannon information and entropy have been applied to a number of information retrieval tasks such as to formalize the probabilistic model, to design practical retrieval systems, to cluster documents, and to model texture in image retrieval. In this report, the concept of entropy is used for a different purpose. It is shown that any positive Retrieval Status Value (RSV)-based retrieval system may be conceived as a special probability space in which the amount of the associated Shannon information is being reduced; in this view, the retrieval system is referred to as Uncertainty Decreasing Operation (UDO). The concept of UDO is then proposed as a theoretical background for term and query discrimination power, and it is applied to the computation of term and query discrimination values in the vector space retrieval model. Experimental evidence is given as regards such computation; the results obtained compare well to those obtained using vector-based calculation of term discrimination values. The UDO-based computation, however, presents advantages over the vector-based calculation: It is faster, easier to assess and handle in practice, and its application is not restricted to the vector space model. Based on the ADI test collection, it is shown that the UDO-based Term Discrimination Value (TDV) weighting scheme yields better retrieval effectiveness than using the vector-based TDV weighting scheme. Also, experimental evidence is given to the intuition that the choice of an appropriate weighting scheme and similarity measure depends on collection properties, and thus the UDO approach may be used as a theoretical basis for this intuition.

Introduction

Shannon (1948) defined the concept of information as one's freedom of choice (to select from alternatives). He also introduced a measure for information, which has maximum value when one has total freedom of choice (i.e., all the alternatives can be selected with equal probability), and has minimal value when one is constrained in selection (i.e., one is "forced" to select exactly one alternative). The formula proposed as a measure for information is an expression of the quantity of Shannon information, and is that of an entropy. While entropy is interpreted as a measure of uncertainty, information is viewed as a reduction in the level of uncertainty. Thus, the amount of Shannon information may also be viewed as a certain expression of uncertainty level; they can, in principle, be used as equivalent concepts. When it is known exactly what to select, then uncertainty is decreased, but when we are free to choose any alternative we want, then uncertainty is highest.

The concept of and formula for entropy has been used in Information Retrieval (IR) in a number of ways.

As early as 1969 (Meetham, 1969), and somewhat later in 1977 (Guazzo, 1977), the concepts of entropy and Shannon information have been applied to IR evaluation as better alternatives to precision and recall. In the 1980s, the maximum entropy principle (MEP) was applied to IR (Cooper & Huizinga, 1982; Kantor, 1984). Formally, MEP can be expressed as a constrained optimization problem, in which one wishes to determine the probability distribution associated to a random variable over a discrete space that has the greatest entropy subject to constraints (these express the knowledge that we impose upon this distribution). In IR, MEP can be formulated as follows. Let the observation of a document with respect to a given query be an elementary event ω. The probability $p(\omega)$ of an event depends on whether the document is relevant or not, and whether it contains query terms or not. The retrieval system aims at maximizing the associated entropy $\Sigma_{\omega} p(\omega)\log p(\omega)$ subject to constraints (such as, e.g., the probabilities of relevant/nonrelevant documents to contain query terms). MEP proves useful as a formal research tool; Greiff and Ponte (2000) show that MEP can be applied as a formal framework in which the probabilistic IR model (Robertson &

Received December 4, 2002; revised November 7, 2003; accepted November 19, 2003

© 2004 Wiley Periodicals, Inc. ● Published online 5 February 2004 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/asi.20008

Sparck Jones, 1977) can be obtained. At the same time, it seems to be less effective when applied to retrieval in practice: extensive experiments show that MEP works well for small document collections but seems to be progressively worse for larger ones (Kantor & Lee, 1998). However, MEP proved useful in text classification tasks as shown by experiments carried out in Nigam, Lafferty, and McCallum (1999). Entropy has been applied to other IR tasks as well. Let C denote a cluster of documents, and $p(C)$ the probability that a relevant document belongs to cluster C (i.e., the ratio between the number of relevant documents in cluster C and the total number of documents in that cluster). Then, the associated entropy can be used as a measure of the clustering process (Fujii & Ishikawa, 2001). Let $p(y)$ denote the probability of a word y as its frequency (i.e., its count over total number of words). Then, a measure of the reduction in uncertainty about whether the word y will be the next word in a sequence of text (given that x was the previous word) can be expressed by entropy (Berger & Lafferty, 1999). Entropy is used for texture modeling in image indexing and retrieval (Yoo, Jang, Jung, Park, & Song, 2002). The information content of a collection of documents consisting of, not necessarily disjoint, classes is the entropy associated to class cardinalities, which is being reduced, in the retrieval process, from its maximum (Baclawski & Simovici, 1996). McClean and Ding (2000) have examined the performance, in terms of precision, of the vector space model of IR when different weighting schemes are being used, and concluded that the entropy weighting scheme is the best global one. Tan, Wang, and Lee (2002) have used MEP for text categorization, and showed that the use of bigrams in addition to single words can increase performance.

Hence, it can be seen that entropy (Shannon information) has been used to formalize the probabilistic IR model, construct practical retrieval systems, cluster documents, and model texture in image retrieval. The present report aims at applying it for a different purpose, namely: it will be shown that a retrieval system using positive Retrieval Status Value (RSV) for retrieval may be conceived as a probability space in which the quantity of the associated amount of Shannon information is being reduced. This result will be applied to the calculation of term and query discrimination values in the vector space model.

Probability Space and Amount of Shannon Information

In order to fix the ideas, the concept of a probability space is briefly recalled (Kolmogoroff, 1933). This will be followed by a short review of the concept of Shannon (1948) information and its measure.

Probability Space

Given a set Ω called *universe*, let its elements be called elementary *events*. A set $\mathcal{T} \subseteq \mathcal{P}(\Omega)$ is called a σ-algebra if $\Omega \in \mathcal{T}$, and

$$A \cap B \in \mathcal{T}, A \cup B \in \mathcal{T}, \Omega \backslash A \in \mathcal{T}, \ \forall A, B \in \mathcal{T} \quad (1)$$

A *probability measure* is a function P: $\mathcal{T} \rightarrow [0; 1]$ satisfying the following properties:

$$P(\Omega) = 1; A \cap B = \varnothing \Rightarrow P(A \cup B) = P(A) + P(B); \ \forall A, B \in \mathcal{T} \quad (2)$$

The triple $(\Omega, \mathcal{T}, P) = \Psi$ is called a *probability space*.

Amount of Shannon Information

Given events (alternatives) E_j, $j = 1, \ldots, m \in \mathbf{IN}$ (**IN** denotes the set of natural numbers); let p_j denote the probability to select alternative (probability of occurrence of event) E_j. Information is conceived as being different from meaning, and defined as one's freedom of choice. A measure H for information is defined as follows:

$$H = -k \sum_{j=1}^{m} p_j \log_2 p_j \quad (3)$$

where k is a positive constant, which amounts to a choice of a unit of measure; in what follows, k will taken as being equal to 1. The quantity H satisfies the following properties:

1. The amount of information is zero if and only if exactly one alternative is selected:

$$\lim_{\substack{p_k \to 1 \\ p_j \to 0, \forall j \neq k}} H = 0 \Leftrightarrow (p_k = 1; p_j = 0, \ \forall j \neq k) \quad (4)$$

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

$$\left(p_j = \frac{1}{m}, j = 1, \ldots, m\right) \Rightarrow H = \max H = \log_2 m \quad (5)$$

3. The farther apart p_j from each other the smaller the amount of information:

$$P < P' \Rightarrow H > H' \text{ where } P = \sum_{j=1}^{m} \left(\frac{1}{m} - p_j\right)^2 \quad (6)$$

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 (i.e., 1 bit). Although it is not stated explicitly, from properties (4)–(6), it follows that $\sum_{j=1}^{m} p_j = 1$; otherwise H could, for example, be null for $p_j = 1$, $j = 1, , m$, too (which is mathematically true), and the, mathematically correct, maximum of H would be reached for $p_j = 2^{-(1/\ln 2)} = 0.368$, $j = 1, \ldots, m$ (all partial derivates are zero).

The property (6) will play an important role in the present report; 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 $p \cdot \log(1/p) > (q - a)\log(1/(q - a))$; whereas 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 while the others closer to zero) the amount of information (or uncertainty) becomes smaller, i.e., the freedom to select becomes more restricted. This entitles us to introduce the following.

Definition 1. An operation (procedure, process, mechanism) that spreads the probabilities p_j from $1/m$ is called an *uncertainty decreasing operation* (UDO).◆

Thus, any operation that constrains the freedom (and thus reduces the uncertainty) to select is a UDO.

Example 1. Numerical minimization of a function based on the gradient method: the freedom to select a direction to follow is decreased because only that given by the gradient can be followed. 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. 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.

Retrieval Status Value-Based Retrieval as Uncertainty Decreasing Operation

Based on the concepts of a probability space and UDO, it will be shown that any retrieval model or system based on positive RSV, e.g., vector space, probabilistic, Boolean, coordination level matching, fuzzy, connectionist interaction, link analysis retrieval models (Dominich, 2001), may be conceived as a probability space that decreases the amount of the associated Shannon information, i.e., it is a UDO probability space.

We first prove that a probability space having its probability measure defined in a certain way (normalized so that its values sum up to unity) is a UDO.

LEMMA. Let $\Psi = (\Omega, \mathcal{T}, P)$, $|\Omega| = m$, denote a probability space with the probability measure P defined as follows:

$$P(X) = \begin{cases} \dfrac{\rho_j}{\sum_{k=1}^{m}\rho_k}, \; X = X_j \in \Omega \\ \\ P'(X), \text{ otherwise} \end{cases} \tag{7}$$

where not all p_{j} are equal to each other, i.e., $\exists\, k \neq s$ such that $p_{\mathrm{k}} \neq p_{\mathrm{s}}$. (An explicit formula for P' does not play any role in this context.) Then the probability space Ψ is a UDO.

Proof. The space Ψ is a UDO if it spreads the probabilities from $1/m$ (Def. 1), i.e., we have to show that

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

We can write

$$P = \sum_{j=1}^{m}\left(\frac{1}{m} - \frac{\rho_j}{\sum_{k=1}^{m}\rho_k}\right)^2 = \frac{\sum_{j=1}^{m}\rho_j^2}{(\sum_{k=1}^{m}\rho_k)^2} - \frac{1}{m} > 0,$$

which is equivalent to the following inequality:

$$m\sum_{j=1}^{m}\rho_j^2 > \left(\sum_{k=1}^{m}\rho_k\right)^2.$$

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

$$m\sum_{j=1}^{m}\rho_j^2 - \left(\sum_{k=1}^{m}\rho_k\right)^2 = \sum_{\substack{k=1\\ s=k+1}}^{\substack{m\\ m-1}} (\rho_k - \rho_s)^2 > 0.◆$$

It can now be shown that any RSV-based retrieval system can be conceived as a probability space that decreases the amount of information.

Theorem 2. Any positive RSV-based retrieval system is a UDO probability space Ψ.

Proof. Given an RSV-based retrieval system. Let ρ_j denote the RSV of document D_j relative to query Q. In other words, ρ_j may be viewed as representing a degree of the choice of document D_j as a response to query Q. The higher the value of ρ_j, the higher the chance of document Dj to be selected as an answer. A sequence $\langle\boldsymbol{\rho}\rangle = \rho_j, \ldots, \rho_m$ can be defined, which represents the choices of all documents relative to query Q. (The no-hit case, i.e., when all the ρ_j are null, can be excluded as trivial.) Using the sequence $\langle\boldsymbol{\rho}\rangle$, a sequence $\langle\boldsymbol{P}\rangle = p_1, , p_m$, where

$$p_j = \rho_j/\sum_{k=1}^{m}\rho_k, \; j = 1, \ldots, m$$

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

$$\Psi = (\Omega, \mathcal{T}, P), \quad \Omega = \{D_1, \ldots, D_m\},$$

$$P(D_j) = p_j, \quad P(X) = P'(X) \text{ if } X \neq D_j.$$

Hence (Lemma), the amount

$$H = -\sum_{j=1}^{m} p_j \log_2 p_j \tag{8}$$

of the Shannon information corresponding to the retrieval situation as a whole is decreased, and thus the RSV-based retrieval system is a UDO.◆

Example 2. 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. If the retrieval function ρ is the dot product, then the chances ρ_1, ρ_2, and ρ_3 are

$$W^{\mathrm{T}}\mathbf{q} = \begin{bmatrix} 2 & 0 \\ 1 & 3 \\ 1 & 2 \end{bmatrix} \times \begin{bmatrix} 0 \\ 1 \end{bmatrix} = \begin{bmatrix} 0 \\ 3 \\ 2 \end{bmatrix},$$

while 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 $\log_2 3 = 1.585$. If the retrieval function is the Cosine measure and the weights are max normalized, 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 $\log_2 3 = 1.585$.

UDO-Based Study of the Vector Space Model

In this section, the concept of UDO will be applied to the study of term and query discrimination power in the vector space model of IR.

Vector Space Model

In order to fix the ideas, the vector space model is recalled briefly. Given a set $T = \{t_i | i = 1, \ldots, n \in \mathbf{IN}\}$ of elements called *terms* (all terms are different, so they may form a set), and entities D_j, $j = 1, \ldots, m \in \mathbf{IN}$ called *documents* (there may be identical documents, so they may not always form a set). Every document is associated, conceived as or characterized by, a finite sequence of terms, i.e., $D_j \sim \langle t \rangle_j = t_{1j}, \ldots, t_{kj}, \ldots, t_{p_j j}$. Let f_{kj} denote the multiplicity of occurrence of term t_k in document D_j. Then, the following set notation can be used too $D_j \sim \{(t_k, f_{kj}) | t_k$ belongs to $\langle t \rangle_j\}$. Let $w_{kj} \in \mathbf{IR}_+$ ($\mathbf{IR}_+$ denotes the set of positive real numbers) denote the *weight* of term t_k for document D_j, i.e., a numerical measure of the degree to which a term pertains to or characterizes a document. Thus, a matrix $W = (w_{kj})_{n \times m}$, 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 following weighting schemes are generally accepted and widely used in practice (Belew, 2000; Berry and Browne, 1999):

$$\text{Term occurrence:} \quad w_{kj} = f_{kj} \tag{9}$$

$$\text{Normalised term frequency:} \quad w_{kj} = \frac{f_{kj}}{\sqrt{\sum_{k=1 \ldots n} (f_{kj})^2}} \tag{10}$$

Normalised inverse document frequency:

$$w_{kj} = \frac{f_{kj} \cdot \log \dfrac{m}{F_k}}{\sqrt{\sum_{k=1 \ldots n} \left(f_{kj} \cdot \log \dfrac{m}{F_k} \right)^2}} \tag{11}$$

where F_k denotes the number of documents in which the term t_k occurs. Thus, every document D_j is represented by a vector $\mathbf{w}_j = (w_{1j}, \ldots, w_{nj}) \in \mathbf{IR}^n$ of weights. Let Q denote a *query*, and $\mathbf{q} = (q_1, \ldots, q_n)$ the corresponding query vector of weights. 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 the following are generally accepted and widely used in practice (Meadow, Boyce, & Kraft, 1999):

$$\text{Dot product:} \quad \rho(\mathbf{w}_j, \mathbf{q}) = \sum_{k=1}^{n} w_{kj} q_k \tag{12}$$

Cosine measure: $$\rho(\mathbf{w}_j, \mathbf{q}) = \frac{\sum_{k=1}^{n} w_{kj} q_k}{\sqrt{\sum_{k=1}^{n} w_{kj}^2 \cdot \sum_{k=1}^{n} q_k^2}} \quad (13)$$

Dice's coefficient: $$\rho(\mathbf{w}_j, \mathbf{q}) = \frac{\sum_{k=1}^{n} w_{kj} q_k}{\sum_{k=1}^{n} (w_{kj} + q_k)} \quad (14)$$

UDO as an Entropy-Based Theory for Term Discrimination and Its Computation

The Term Discrimination Model (TDM) was introduced in Salton, Yang, and Yu, (1974, 1975) as a contribution to the automatic indexing theory in the vector space model of information retrieval. The TDM is based on the underlying assumption that a "good" term causes the greatest possible separation of documents in the vector space, whereas a "poor" term makes it difficult to distinguish one document from another. Each term under focus is assigned a Term Discrimination Value (TDV) defined as the difference between space "densities" before and after removing that term. The space "density" Δ is defined as the average pairwise similarity ρ between documents:

$$\Delta = \frac{1}{m(m-1)} \sum_{\substack{i,j=1 \\ i \neq j}}^{m} \rho(D_i, D_j) \quad (15)$$

Alternatively, space density Δ can be computed, faster, as the average similarity between documents and a centroid document (defined as one in which the terms have frequencies equal to their average frequencies across the collection of documents). Let Δ_{bk} and Δ_{ak} denote the space "densities" before and after removing term t_k, respectively; then the TDV$_k$ of term t_k is defined as follows:

$$\text{TDV}_k = \Delta_{bk} - \Delta_{ak} \quad (16)$$

The best discriminators generally have positive TDVs, whereas the worst discriminators usually have negative TDVs. Terms having TDVs around zero do not modify the space density considerably when used as index terms. The TDV can be used for the following purposes:

1. To decide which terms should be used as index terms (Yu & Salton, 1977): terms with average document frequencies (between approximately $m/100$ and $m/10$) usually have positive TDVs, and can be used directly for indexing purposes; terms whose document frequency is too high generally have negative TDVs, and are the worst discriminators; too rare or specific terms have TDVs near zero, and should not be used directly as index terms.
2. Weights computation for terms (Salton, 1986): while the Inverse Document Frequency (IDF) method prefers low-frequency terms, they are not preferred in the TDM; thus, in the TDM, the weight w_{kj} of term t_k in document D_j should be computed as $w_{kj} = f_{kj} \cdot \text{TDV}_k$ (instead of $w_{kj} = f_{kj} \cdot \text{IDF}_k$).
3. Thesaurus construction (Crouch & Yang, 1992): TDVs of terms are used to construct thesaurus classes.

Practical research carried out in TDM has since highlighted several insights as follows. Crouch and Yang (1992) and Dubin (1995) have found that there is not a direct and exact correlation between discrimination value on the one hand, and documents and term frequency on the other hand. The computation of the TDV (formula 16) is expensive, which hinders its practical application. Willett (1985) developed a faster method, and also showed that whether a term is a poor or good discriminator depends on the similarity measure used: the dot product yields a monotonically decreasing relationship between TDV and document frequency, the Euclidean distance leads to an increasing relationship. As shown also in Dubin (1995), the TDV depends not only on the similarity measure used but also on the weighting method used, and stop list. In light of these results, it turns out that there are different correlations between TDV, weighting schemes, and similarity measures.

The underlying assumption of TDM is based on a topological view: separation, distinguishable, density, sparsity. This view is being applied to the vector space model whose typical space is the n-dimensional orthonormal linear space, i.e., each dimension corresponds to a term, the documents are represented as vectors of weights, the fact that the terms are considered to be independent of one another is modeled by the pairwise perpendicular coordinate axes, the usual similarity measures (Dot product, Cosine measure, Dice's coefficient, Euclidean distance) make sense because the space is Euclidean. Thus, the TDM as a theory and the TDV as a computation method cannot be applied to other, aside from the vector space model, RSV-based information retrieval models that do not use linear or Euclidean space, and hence whose similarity is not based on inner, or dot, product. However, both from a practical and theoretical point of view, it would be useful to have a means to compute terms TDV in any RSV-based retrieval model.

The concept of UDO introduced in the present report is valid for any RSV-based retrieval model; it does not necessarily assume the existence of a linear space. In what follows, it will be proposed that the UDO view of RSV-based retrieval models be applied to the study and computation of "discriminatory" or "separation" power of individ-

TABLE 1. Statistics for the ADI test collection used in the UDO-based TDV computation.

Subject area	Information science
Type	Homogeneous
No. of documents	82
No. of terms	915

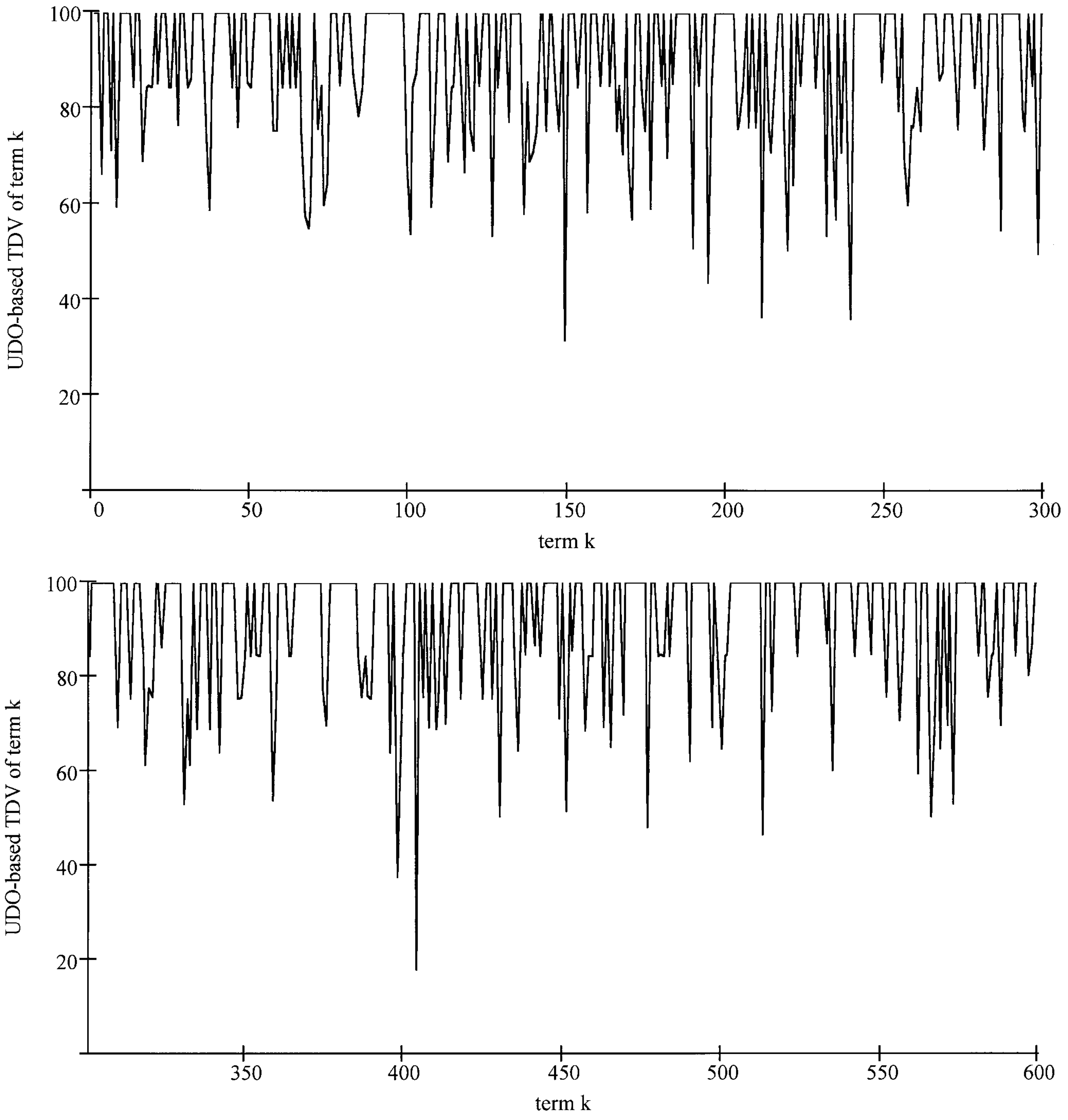

FIG. 1. UDO-based TDV for the ADI test collection. Discrimination values of terms are computed as entropy reductions. On the horizontal axis, term k means term t_k, i.e., the kth term in the list of index terms. On the vertical axis, the corresponding TDV is shown as the percentage of entropy reduction from the maximum entropy value. The Cosine similarity measure and the normalized inverse document frequency weighting scheme were used. For example, the 150th term, i.e., $k = 150$, reduces entropy by 30%.

ual terms. Let Q_k denote a single term query containing exactly one term, t_k, where the term t_k is selected from a list of index terms. Then the corresponding entropy H_k (computed using formula 8) will be a measure of the extent to which the term t_k is able to reduce the retrieval system's uncertainty in selecting documents (for returning answers). Thus, in the UDO view, the TDV of a term is based on how much it reduces this entropy, associated to a probability space, rather than how much it reduces space density in Euclidean, and hence topological, space.

In what follows, experimental evidence will be shown, using the ADI test collection, for the computation of UDO-based TDV, and how it compares to vector-based TDV. Table 1 shows the ADI statistics for the experiments. The terms were selected from the ADI documents; they were TIME stop listed and Porter stemmed.

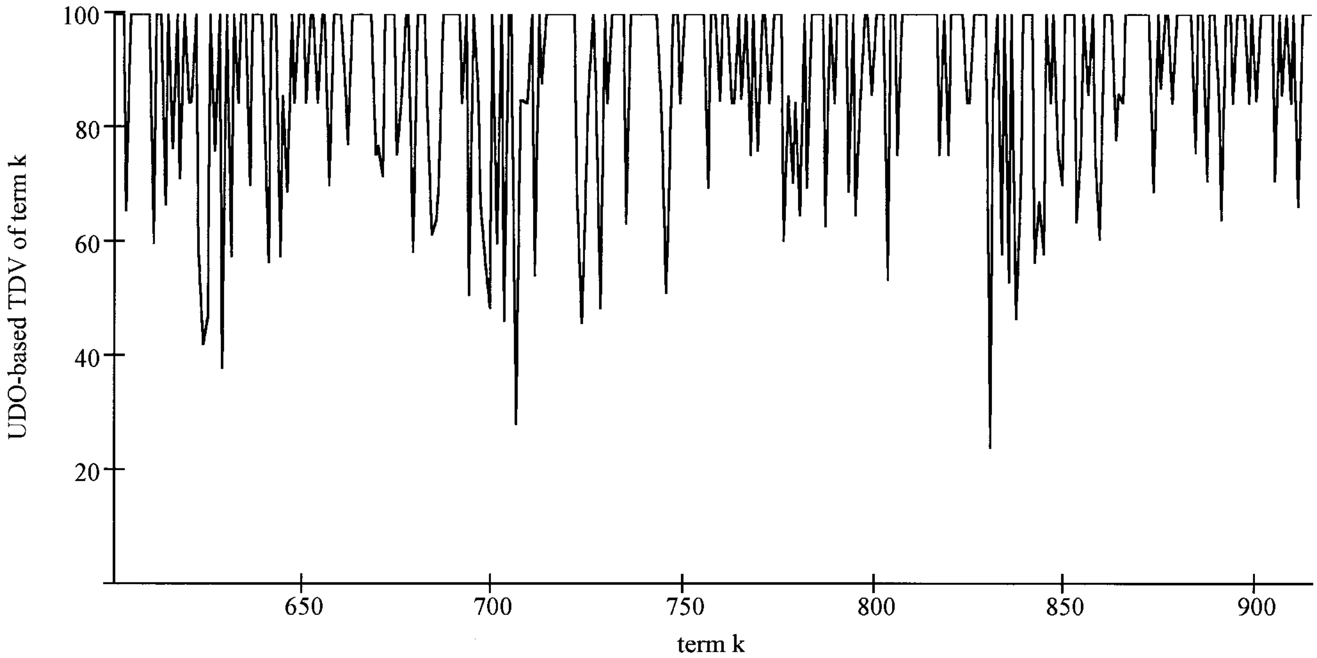

FIG 1. (Continued).

Figure 1 shows the UDO-based TDVs of terms computed as entropy reductions (i.e., H_{max} H in %). Figure 2 shows the vector-based TDVs calculated as space density variations (formula 16) using centroid document. In both cases, the similarity measure used was the Cosine measure (formula 13), and the weights were computed using the normalized inverse document frequency scheme (formula 11). Comparing the UDO-based TDVs and the vector-based TDVs, and taking into account term frequencies as well (Fig. 3), one can state the following.

About half the terms have their UDO-based TDV equal to 100%, i.e., they practically totally reduce entropy. The frequency of each such term is equal to 1, i.e., each such term occurs in exactly one document. The vector-based TDV (Fig. 4) of every such term is positive, with a mean value of 6.4×10^{-6} and standard deviation of 0. These terms are good discriminators.

Figure 5 shows the vector-based TDV of terms whose UDO-based TDV is <100% and >80%. The mean of vector-based TDV is equal to 6.056×10^{-6} having zero standard deviation. The document frequency of every such term is 2. Most of these terms have positive vector-based TDV.

Figure 6 shows the vector-based TDV of those term whose UDO-based TDV belongs to the interval (40, 80%). The vector-based TDV is practically zero for every such term. Document frequency has a mean equal to 5.066 with standard deviation equal to 2.406.

A few terms reduce entropy with <40%. They are poor discriminators. Their document frequency is high, with a mean value of 25 and standard deviation of 8.93. Their vector-based TDV are negative and low values. The dotted line in Figure 7 shows the documents' frequencies of terms (scaled for representation and comparison purposes), whereas the solid line shows the vector-based TDVs. There appears to be a symmetry, and hence a direct proportionality, between document frequency and vector-based TDV in this case.

The results can be summarized as follows:

- the terms that reduce entropy almost entirely, i.e., in the interval (80, 100%), are very good discriminators, and have very low document frequency;
- the terms that hardly reduce entropy, i.e., in the interval (0, 40%), are poor discriminators, and have very high document frequency;
- the terms that reduce entropy in the middle range, in the interval (40, 80%), are indifferent discriminators, and have relatively medium document frequency.

Similar results can be obtained using the vector-based TDV computation; however, the entropy-based TDV calculation method presents the following advantages:

- it is faster than the computation of vector-based TDV with formula 16; the numeric results (i.e., the entropy reduction values given as %) are easier to assess and handle than the vector-based TDV expressed as fractional real numbers with exponent;
- the UDO-based view does not assume the existence of any linear space like the traditional TDM does;
- the UDO-based view can be applied to compute term TDVs in any RSV-based retrieval model (not just the vector space model).

UDO-Based Query Discrimination Power

The concept of UDO introduced in the present report can be used to define and compute a discrimination power for queries, similar to that for terms.

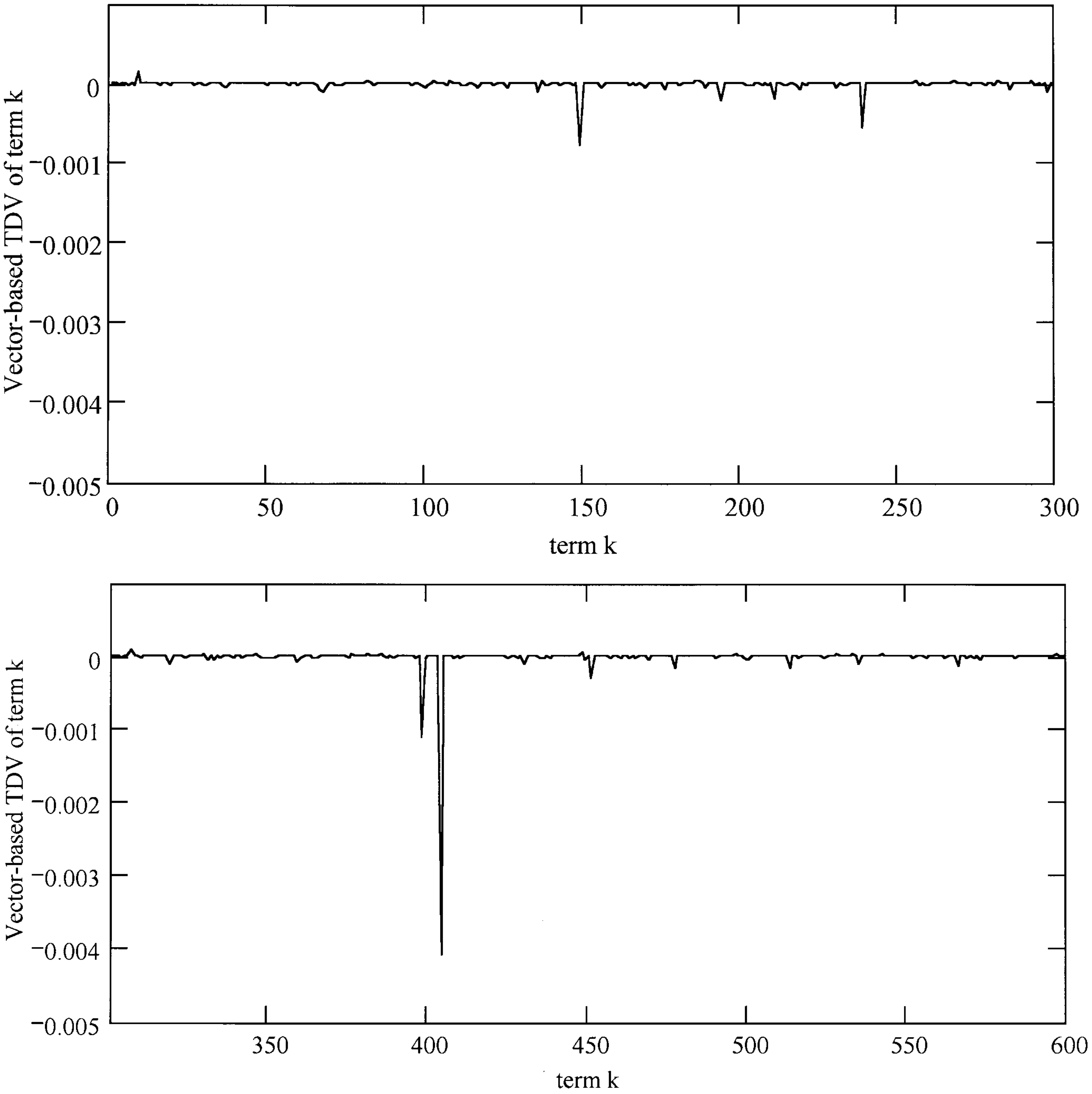

FIG. 2. Vector-based TDV of terms for the ADI test collection. On the horizontal axis, term *k* means term t_k, i.e., the *k*th term in the list of index terms. On the vertical axis, the corresponding TDV is shown computed using centroid document. The Cosine similarity measure and the normalized inverse document frequency weighting scheme were used.

Let *Q* denote a query. Then a discrimination power of query *Q* can be defined as its ability to reduce the uncertainty to select documents to be retrieved, and a measure of this ability will be the quantity H_Q calculated with formula 8; this quantity will be referred to as Query Discrimination Value (QDV).

In what follows, this will be applied to the queries of the ADI test collection. Figure 8 shows the corresponding UDO-based QDVs (solid bars) for all the 35 test queries; the stems represent the number of relevant documents for each query as given in the test collection itself. It can seen that, in general, the fewer relevant answers (shown by stems) a query has the higher its QDV, and vice versa. There is one striking exception: query no. 27; this reads as follows: "Computerized information retrieval systems. Computerized indexing systems." The explanation can be found by examining (Fig. 3) the document frequencies of its terms: "computer" is term number 149, "information" is term number 404, "retrieval" is term number 706, "system" is term number 830, and "indexing" is term number 398. All

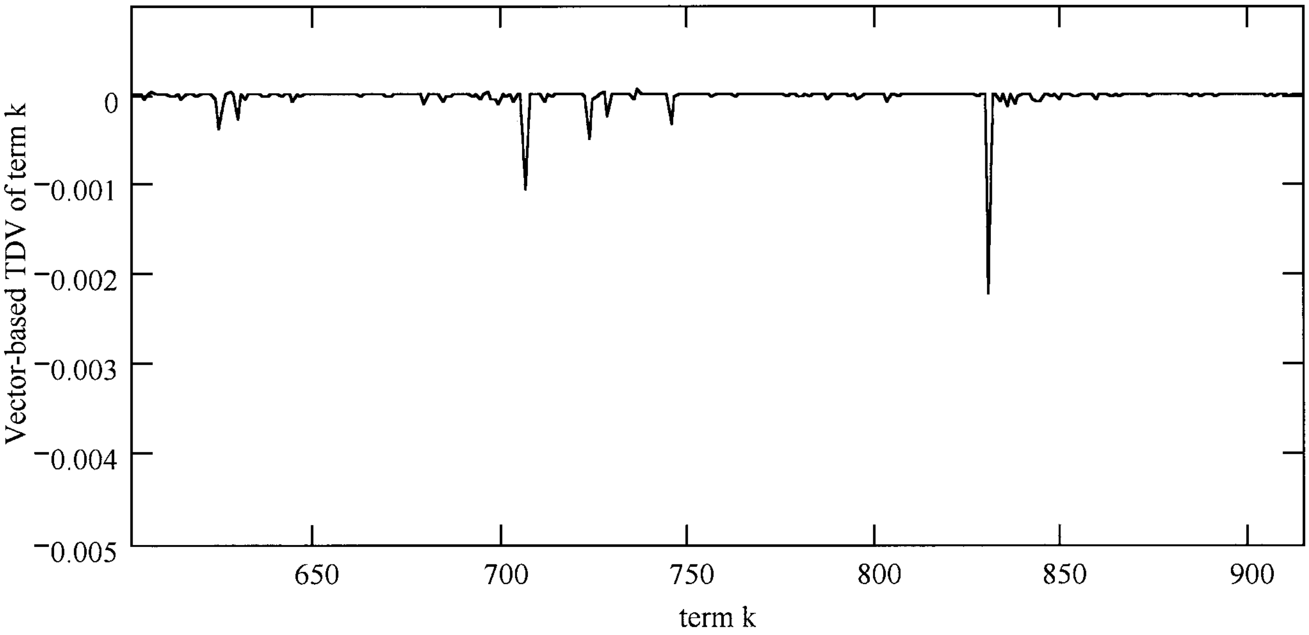

FIG 2. (Continued).

these terms have very high document frequencies, which accounts for this low QDV. In other words, query number 27 is perhaps too general, and maybe should not have been included in the test collection; although it can be used to test recall.

Retrieval Effectiveness Using TDV Weighting Scheme

In order to estimate and test the retrieval effectiveness of a vector space retrieval using a weighting scheme based on UDO-TDV, i.e., as given in UDO in point 2 in Entropy-Based Theory for Term Discrimination and Its Computation:

$$w_{kj} = f_{kj} \cdot \mathrm{TDV}_k,$$

the standard precision-recall measurement was carried out over the ADI test collection using both the UDO-based and vector-based TDVs; thus, the performances can be compared. Figure 9 shows the precision-recall curves obtained. The similarity measure used was the dot product measure in both cases. Precision and recall were computed at the first twenty seen documents. It can be seen that using UDO-based TDVs to enhance the term weights yields increased precision at higher recall levels (above 50%); while both methods perform equally well at lower recall levels. The average precision is equal to 0.450 when using UDO-based TDVs, and to 0.365 when using vector-based TDVs; thus, it can be said that the UDO-based TDV weighting scheme yields higher performance.

Analyzing the details of the usual calculations of precision and recall (not shown here as the process of such computation is well known), we can note the following. The question 27, which was the striking exception in Figure 9, contributed to a fairly great extent to increasing precision at higher recall levels: using UDO-based TDVs, the number of relevant documents retrieved was twice as much as in the case of vector-based TDVs. Neither weighting schemes found any relevant documents to questions 14: from Figure 9, we can see that this question has very few relevant documents (namely: two) but it is a quite discriminatory question (UDO = 43%). Apart from that, in the case of the vector-based TDV weighting scheme, no relevant documents were found for further questions:17, 20, 23, 28, and 32 (the UDO-based weighting scheme did find relevant documents for these questions). Also, we can note that, in general, the number of found relevant documents to every query was higher in the case of the UDO-based TDV weighting scheme than in the case of the vector-based TDV scheme.

Method for the Study of the Relationship Between Entropy Reduction, Weighting Scheme, and Similarity Measure

Based on Theorem 2, the following method can be proposed for the practical study of the relationship between entropy reduction, weighting scheme, and similarity measure in RSV-based retrieval systems.

1. Computation of entropy reduction (fixed query; fixed weight and RSV computation method):
   a. Implement the IR model under focus (using some data set).
   b. Formulate a query Q.
   c. Compute entropy H (formula 8),
   d. Compute the maximum value of entropy as H_{max} = log m (formula 5).
   e. Calculate entropy reduction as $H_{max} - H$ (alternatively as %).

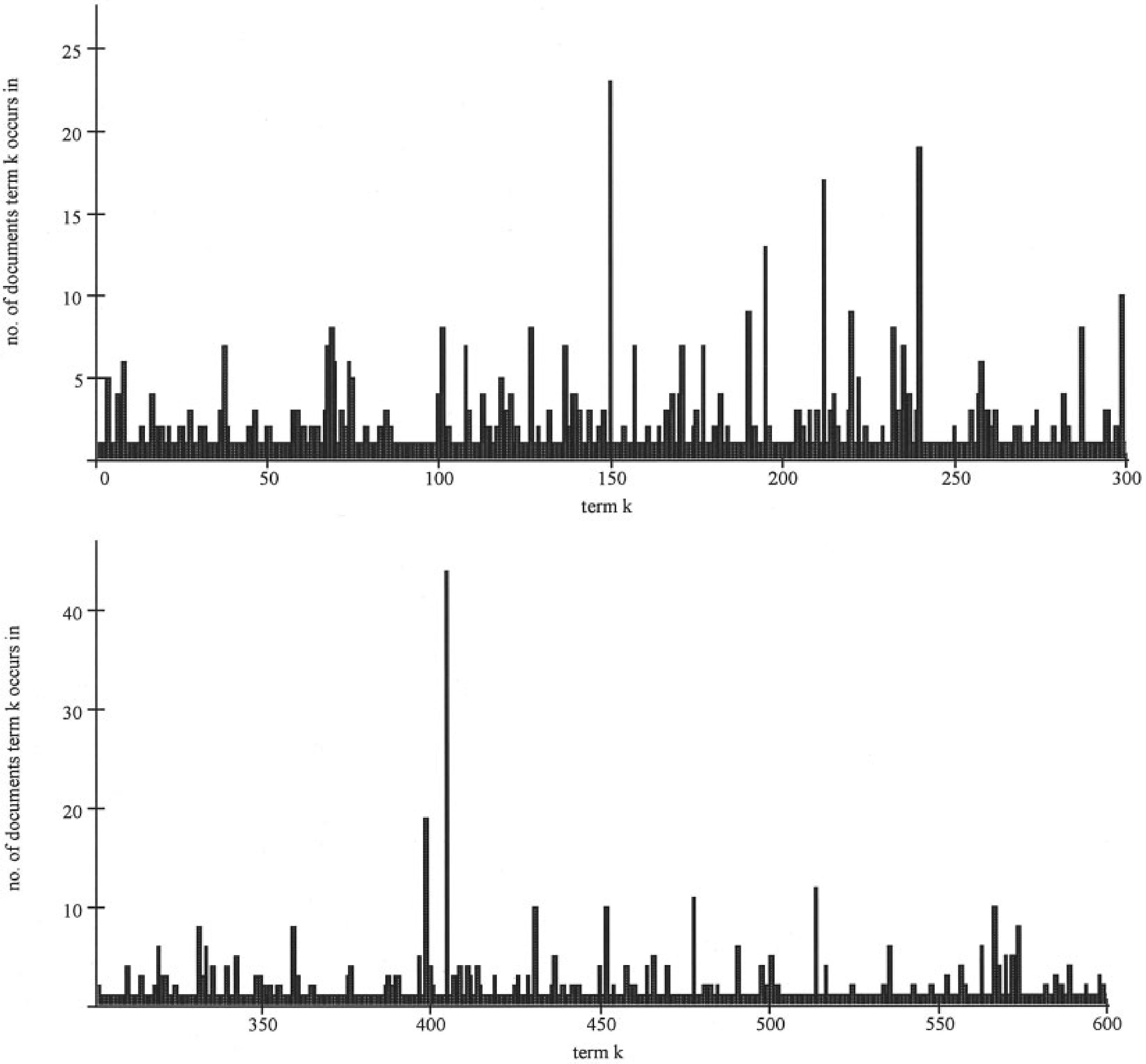

FIG. 3. Document frequency of terms in the ADI test collection. On the horizontal axis, term *k* means the *k*th term in the list of index terms. On the vertical axis, the document frequency of term *k* is shown, i.e., the number of documents in which the term *k* occurs.

2. Estimation of entropy reduction (over several queries):
   a. Formulate (generate) several queries.
   b. Perform steps 1a–e for each query.
   c. Compute the average of the entropy reductions obtained at step 2b.
3. Study of RSV-based IR model:
   a. Implement the IR model under focus (over some data set) using different weighting schemes and RSV computation methods.
   b. Perform steps 2a–c for each weighting and RSV computation method.
   c. Create a table with the results obtained.
   d. Draw conclusions.

Experiments were carried out with three text collections. 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 are as usual).

The third collection of texts, referred to as BELIEFS, contained 2,704 Hungarian belief texts. The choice of this collection was motivated by two reasons: it was not a standard one, and it was special in that 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 fejkõtô alá, az atya, fiú,*

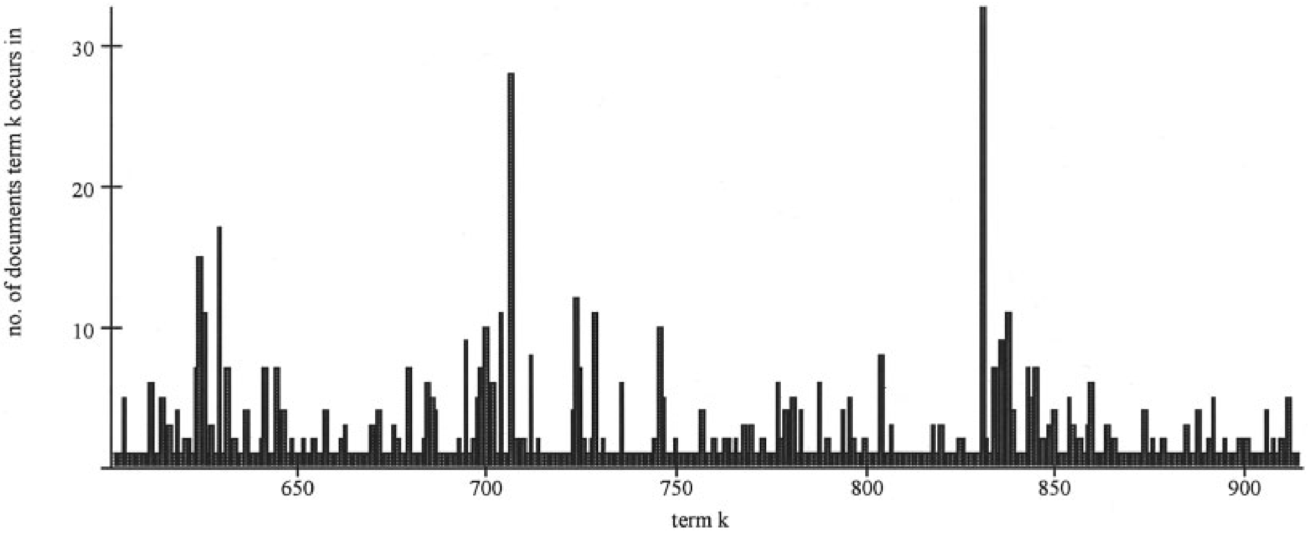

FIG 3. (Continued).

szentlélek nevében. Amen.; and a mixture of older Hungarian spelling and dialect; for example, *Ha a tehenet merrontya a boszorkány, vësznek egy új félliteres 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 were 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, the personal pronoun *aki*, meaning "who," has many different declined forms such as: *aki, akié, akiébe, akiért, akihez, akijé, akik, akiknek, akin, akinek, akinél, akire, akiröl, akit, akitöl, 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 following declined word forms: *csont, csont-*

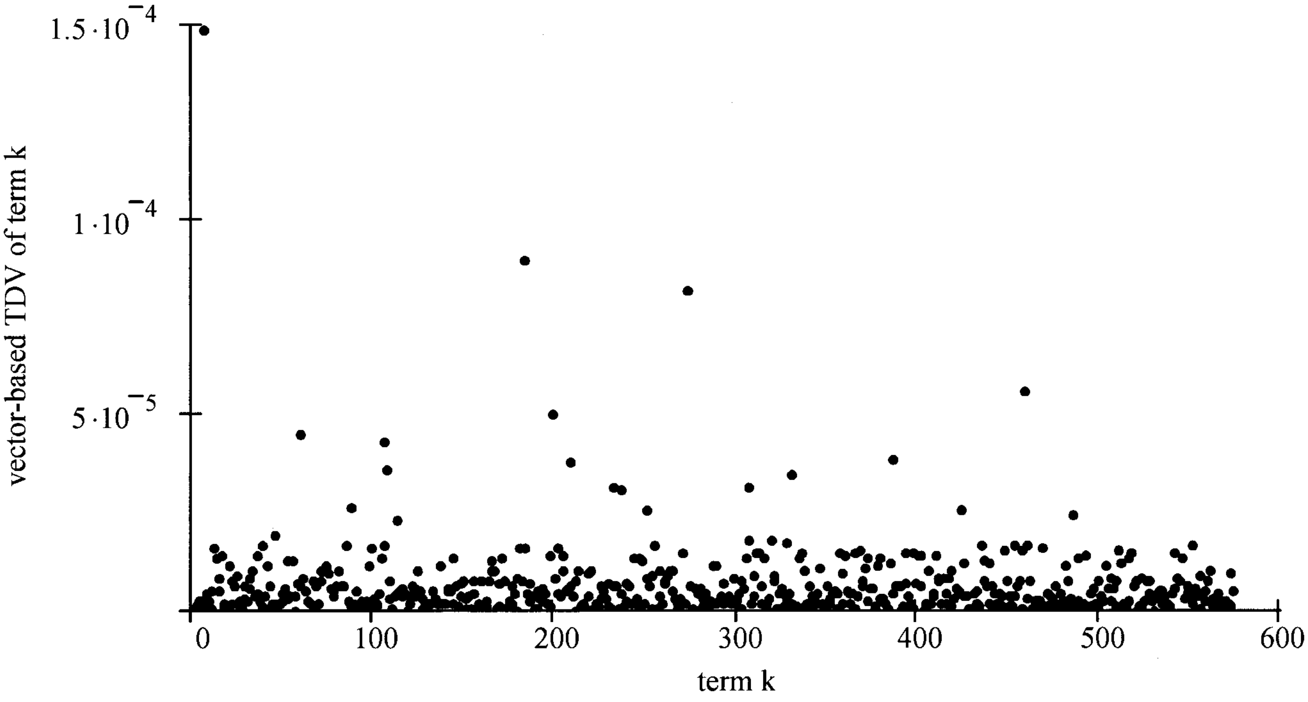

FIG. 4. Vector-based TDV of those terms whose UDO-based TDV is 100% (ADI test collection). The document frequency of every term is equal to 1. The mean value of TDVs is equal to 6.4 × 10^{-6} with a standard deviation of 0.

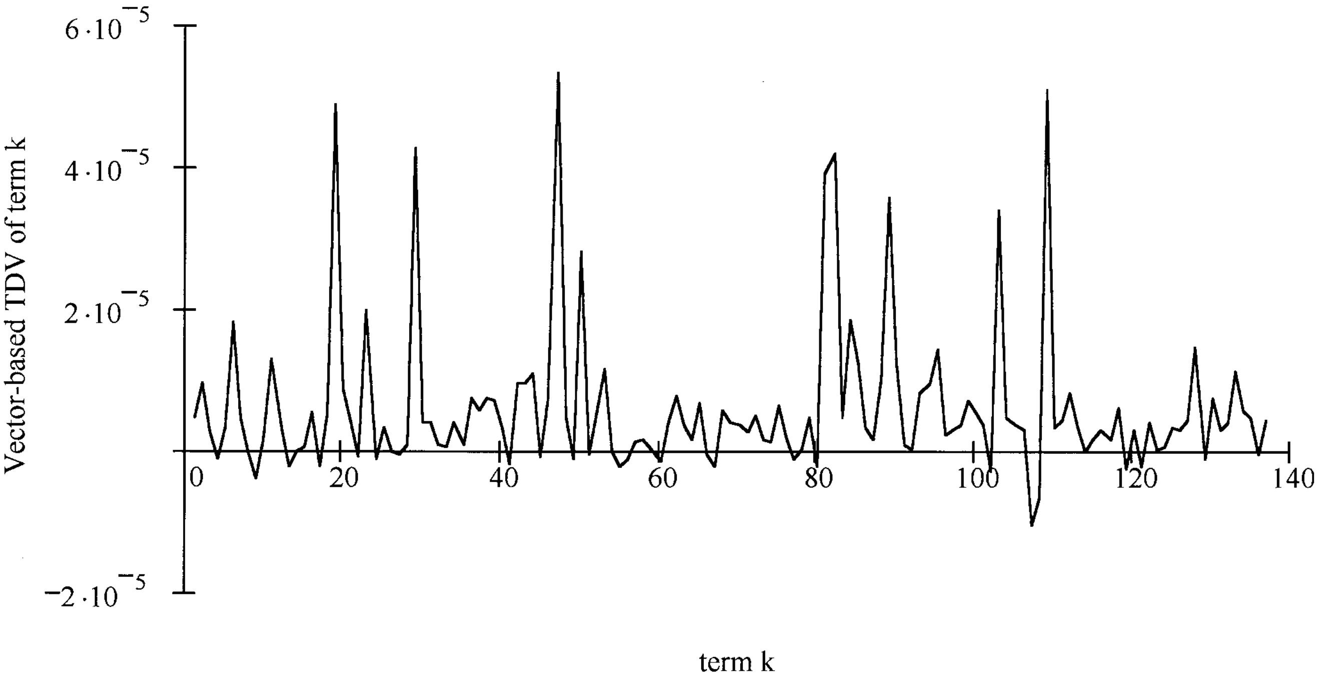

FIG. 5. The vector-based TDV of terms whose UDO-based TDV is <100% and >80%. The mean of vector-based TDV is equal to 6.056 = 10^{-6} having 0 standard deviation. The document frequency of every such term is 2.

jával, 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 or Finnish, for instance). A further and very special difficulty was posed by old homonym words that are not being used anymore in contempo-

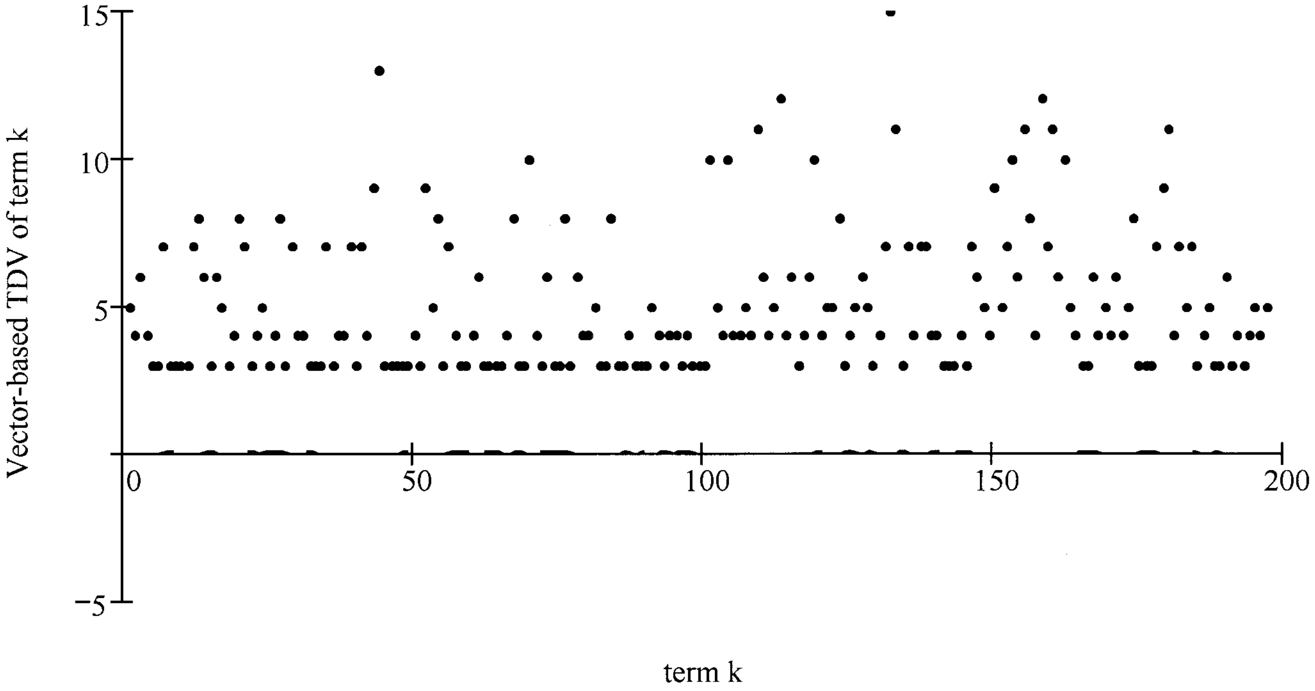

FIG. 6. The vector-based TDV of those term whose UDO-based TDV belongs to the interval (40, 80%). The vector-based TDV (shown as solid line segments along the horizontal axis) is practically zero for every such term. Document frequency (shown as dots) has a mean equal to 5.066 with a standard deviation equal to 2.406.

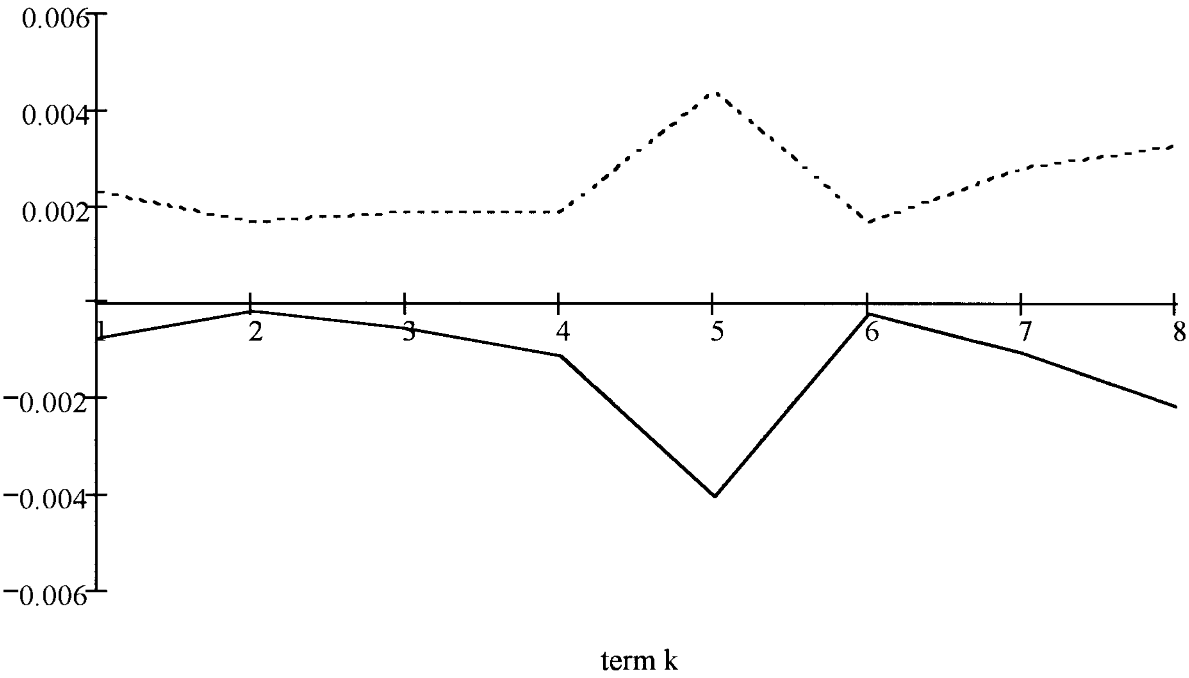

FIG. 7. Vector-based TDV of those terms whose entropy reduction is below 40% (ADI test collection). TDV is shown as a solid line, and the corresponding document frequency as a dotted line.

rary Hungarian; for example, the words *betyöleges, biszbányosok* were replaced by the word *varázs* meaning "magic." The result was a number of 2,607 terms in a correct contemporary Hungarian spelling, which were used as index terms for the belief texts. The average number of terms per text was 15.

Ten—single-term and multiple-term—queries were generated randomly (from the set of index terms to avoid no-hit cases) for each text collection, weighting scheme, and similarity measure separately. The computations were performed following the steps 1–3 above. Two weighting schemes were used for each text collection: term frequency

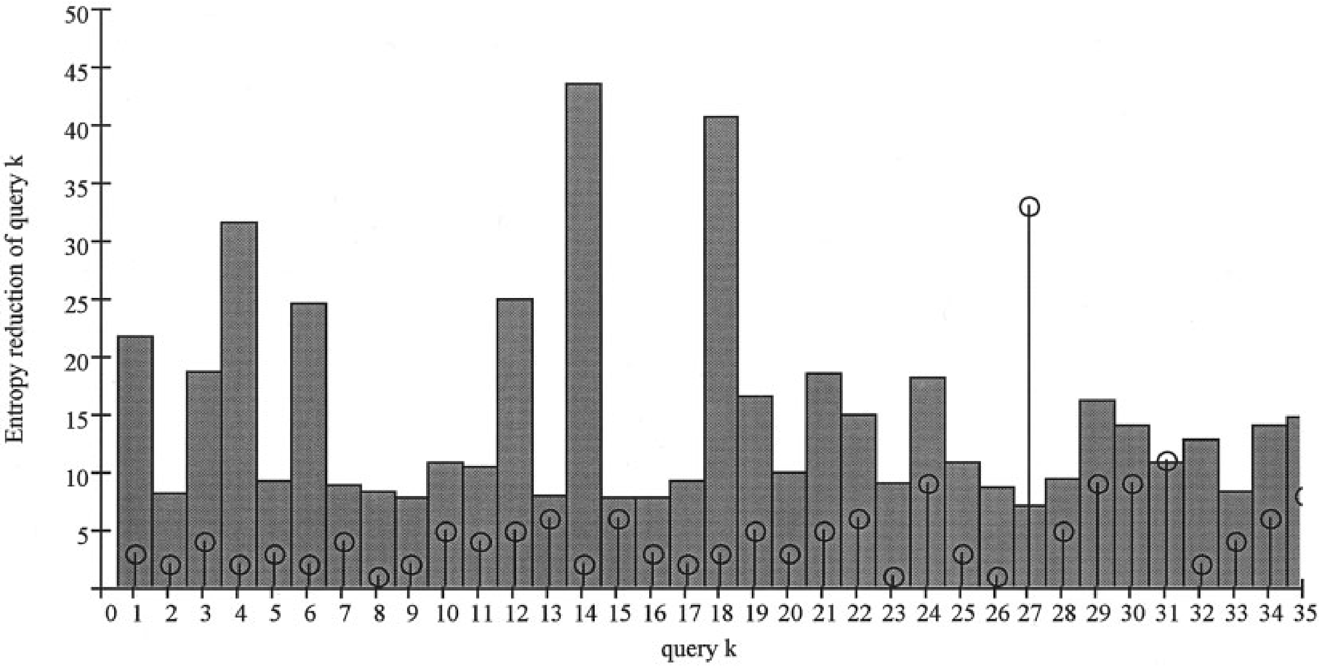

FIG. 8. UDO-based query discrimination values (solid bar) for the ADI test collection. The Dot product similarity measure and the frequency weighting schemes were used. The stems represent the corresponding number of relevant documents for query *k* (horizontal axis) as given in the test collection itself.

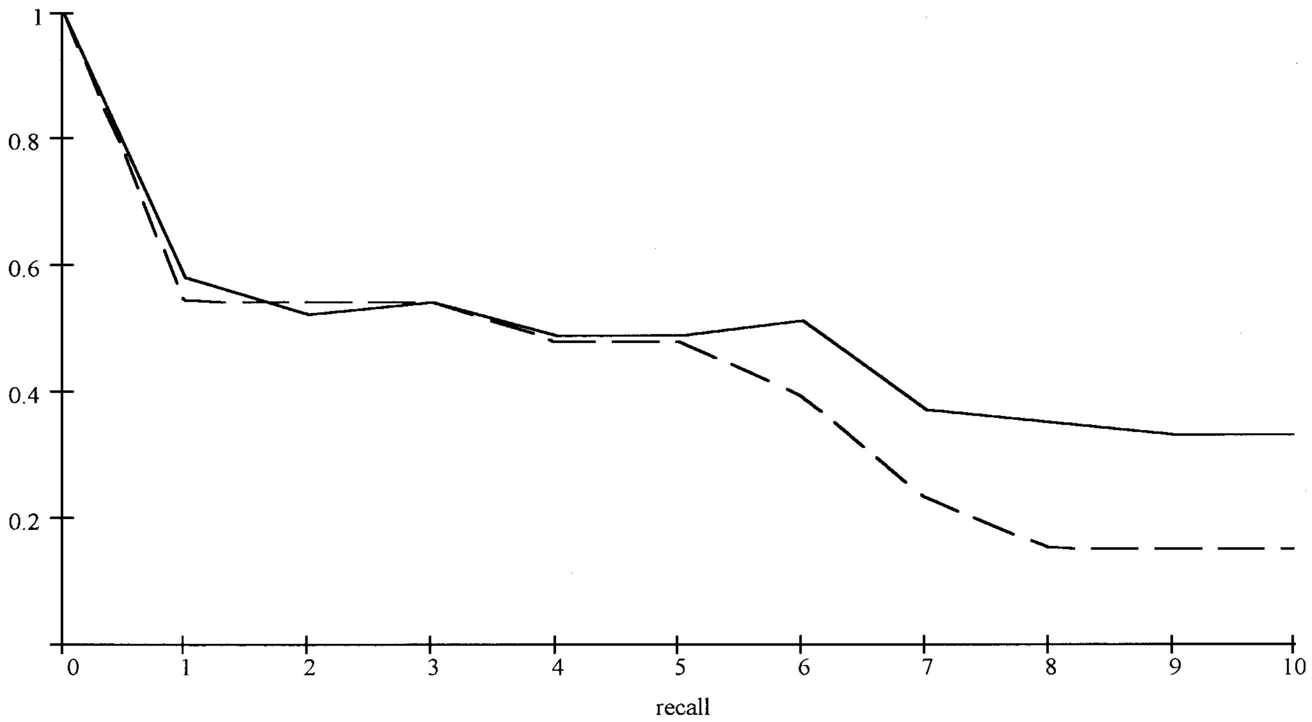

FIG. 9. The 11-point (0, 10,,100%) standard precision-recall curves for the ADI test collection using TDV-based weighting schemes. The weights of the term-document matrix were computed using the TDV weighting scheme $w_{kj} = f_{kj} \cdot \mathrm{TDV}_k$. The solid line is the precision-recall curve corresponding to the UDO-based TDV weighting scheme, while the dashed line is the precision-recall graph corresponds to the vector-based TDV scheme.

(formula 9) and normalized frequency (formula 10), and the three similarity measures given by formulas 12–14. Table 2 shows the results.

The experiments show the following results:

1. The usage of the dot product similarity measure reduces entropy to the greatest extent for every weighting scheme and collection of text.
2. The usage of the normalized frequency weighting scheme reduces entropy to the greatest extent for every similarity measure and collection of text.
3. The usage of the Cosine similarity measure resulted in the least entropy reduction for every weighting scheme and collection of text.
4. The usage of the frequency weighting scheme yielded the least entropy reduction for every similarity measure and collection of text.

These results seem to confirm the intuition that an appropriate weighting scheme should depend on certain properties of the document collection on which the model is to be applied (Berry & Browne, 1999). Thus, for technical or scientific documents, normalized term frequencies are generally recommended, while in the case of more general

TABLE 2. Entropy reduction for several weighting schemes and similarity measures in the vector space model over three different collections.

	Average entropy				
Collection	Cosine	Dice coeff.	Dot prod.	Average entropy	Maximum entropy
Medline					
Frequency	9.943	9.599	9.330	9.624	10.013
Norm. freq.	9.943	8.644	7.520	8.702	
Average entropy	9.943	9.122	8.425		
Time					
Frequency	8.632	8.151	7.809	8.224	8.725
Norm. freq.	8.632	6.184	4.727	6.514	
Average entropy	8.632	7.167	6.268		
Beliefs					
Frequency	11.321	10.959	10.813	10.971	11.401
Norm. freq.	11.321	10.790	10.275	10.665	11.401
Average entropy	11.321	10.874	10.544		

documents (e.g., magazines, encyclopedia), term frequencies may be sufficient. Hence, we may say that the UDO-based approach provides a theoretical basis for the above intuition.

Conclusions

In this report, it was shown that any RSV-based retrieval system may be conceived as a special probability space in which the amount of the associated Shannon information is being reduced; in this view, the retrieval system was referred to as Uncertainty Decreasing Operation (UDO). The concept of UDO was then proposed as a theoretical background for term and query discrimination power, and it was applied to the computation of term and query discrimination values in the vector space retrieval model. Experimental evidence was given as regards such computation. The results obtained compare well to those obtained using vector-based calculation of term discrimination values. The UDO-based computation, however, presents advantages over the vector-based calculation: it is faster, easier to assess and handle in practice, and its application is not restricted to the vector space model. Also, experimental evidence was given to the intuition that the choice of an appropriate weighting scheme and similarity measure depends on collection properties, and thus the UDO approach may be used as a theoretical basis for this intuition. Based on the ADI test collection, it was shown that the UDO-based TDV weighting scheme yields better retrieval effectiveness than using the vector-based TDV weighting scheme.

Acknowledgments

The authors acknowledge the support of the following 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 BELIEFS texts.

References

Baclawski, K., & Simovici, D.A. (1996). A characterization of the information content of a classification. Information Processing Letters, 57, 211–214.

Belew, K.B. (2000). Finding out about. Cambridge: Cambridge University Press.

Berger, A., & Lafferty, J. (1999). Information retrieval as statistical translation. Proceedings of the ACM SIGIR (pp. 222–229).

Berry, M.W., & Browne, M. (1999). Understanding search engines: Mathematical modeling and text retrieval. Philadelphia: SIAM.

Cooper, W.S., & Huizinga, P. (1982). The maximum entropy principle and its application to the design of probabilistic retrieval systems. Information Technology, Research and Development, 1, 99–112.

Crouch, C.J., & Yang, B. (1992). Experiments in automatic statistical thesaurus construction. Proceedings of the 15th Annual International SIGIR Conference (pp. 77–88). Copenhagen, Denmark.

Dominich, S. (2001). Mathematical foundations of information retrieval. Dordrecht: Kluwer Academic Publishers.

Dubin, D. (1995). Document analysis for visualisation. Proceedings of the Annual International ACM SIGIR Conference (pp. 199–204). New York: ACM.

Fujii, A., & Ishikawa, T. (2001). Evaluating multi-lingual information retrieval and clustering at ULIS. Proceedings of NTCIR Meeting on Evaluation of Chinese and Japanese Text Retrieval and Summarization (pp. 250–254). Beijing, China.

Greiff, W.R., & Ponte, J.M. (2000). The maximum entropy approach and probabilistic IR models. ACM Transactions on Information Systems, 18(3), 246–287.

Guazzo, M. (1977). Retrieval performance and information theory. Information Processing and Management, 13, 155–165.

Kantor, P.B. (1984). Maximum entropy and the optimal design of automated information retrieval systems. Information Technology, 3(2), 88–94.

Kantor, P.B., & Lee, J.J. (1998). Testing the maximum entropy principle for information retrieval. Journal of the American Society for Information Science, 46(6), 557–566.

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

McClean, M., & Ding, C.H.Q. (2000). Using advanced mathematics to improve information retrieval precision. (n.d.). Retrieved December 8, 2002, from http://www.lbl.gov/Education/CSEE/cup/Su00/McClean/webpage.html

Meadow, C.T., Boyce, B.R., & Kraft, D.H. (1999). Text information retrieval systems. New York: Academic Press.

Meetham, A.R. (1969). Communication theory and the evaluation of Information retrieval systems. Information Storage and Retrieval, 5, 129–134.

Nigam, K., Lafferty, J., & McCallum, A. (1999). Using maximum entropy for text classification. Proceedings of IJCAI-99 Workshop on Machine Learning for Information Filtering (pp. 61–67). Dortmund, Germany.

Robertson, S.E., & Sparck Jones, K. (1977). Relevance weighting of search terms. Journal of the American Society for Information Science, 27, 129–146.

Salton, G. (1986). Another look at automatic text-retrieval systems. Communications of the ACM, 29(7), 648–656.

Salton, G., Yang, C.S., & Yu, C.T. (1974). Contribution to the theory of indexing. Information Processing 74 (pp. 584–590). Amsterdam: North Holland Publishing Co.

Salton, G., Yang, C.S., & Yu, C.T. (1975). In theory of term importance in automatic text analysis. Journal of the American Society for Information Science, 26(1), 33–44.

Shannon, C. (1948). A mathematical theory of communication. The Bell System Technical Journal, 27, 379–423, 623–656, July, October.

Tan, C.-M., Wang, Y.-F., & Lee, C.-D. (2002). The use of bigrams to enhance text categorization. Information Processing and Management, 38, 529–546.

Willett, P. (1985). An algorithm fro the calculation of exact term discrimination values. Information Processing and Management, 21(3), 225–232.

Yoo, H.-W., Jang, D.-S., Jung, S.-H., Park, J.-H., & Song, K.-S. (2002). Visual information retrieval system via content-based approach. Pattern Recognition, 35(3), 749–769.

Yu, C.T., & Salton, G. (1977). Effective information retrieval using term accuracy. Communications of the ACM, 20(3), 135–142.