Information Processing & Management Vol. 25, No. 1, pp. 39-53, 1989
Printed in Great Britain.

0306-4573/89 $3.00 + .00

III. PROBABLISTIC MODELS

A PROBABILITY DISTRIBUTION MODEL FOR INFORMATION RETRIEVAL

S.K.M. Wong and Y.Y. Yao
Department of Computer Science, University of Regina,
Regina, Saskatchewan S4S 0A2, Canada

Abstract—A probability distribution model for information retrieval is proposed in this article. In contrast to other approaches, each document in the new model is characterized by a simple probability measure on the set of index terms. Based on this interpretation of document representation, two different retrieval strategies are discussed. One is based on utility theory and the other is derived from information theory. The new model not only enhances retrieval effectiveness as demonstrated by experiments, but also provides valuable insight into many fundamental concepts introduced over the years in a variety of retrieval models.

1. INTRODUCTION

There are three major mathematical models in information retrieval: Boolean retrieval, vector space, and probabilistic models.

The Boolean retrieval model is perhaps best understood by virtue of its simplicity and sound theoretical basis. However, in spite of its ability to process structured queries, it has been criticized for its inability to provide ranked output as all retrieved documents are considered equally important. A Boolean request is also apt to retrieve either too many or too few documents. Most of these drawbacks stem from the exact matching strategy adopted by the Boolean retrieval model.

By adopting a partial matching strategy, the vector space model is able to rank documents according to their similarity values with respect to a query. These similarity values are believed to reflect the degree of relevance of each document from the user's point of view. It has been pointed out in [1] that some earlier work with the vector space model did not fully explain the various concepts involved. The misunderstandings of the interactions among these concepts may have led to some inconsistent usage of the model. In fact, some of the fundamental issues have not yet been completely resolved. One of the main criticisms of the standard vector space model is the term pairwise orthogonality assumption. Some attempts [2–5] have been made to remove such a strict assumption but there is still a lack of vigorous justifications for using *linear* similarity functions in these approaches. The necessary and sufficient conditions [6] based on measurement theory that justify the usage of linear discriminant functions have been presented only recently. Nevertheless, the vector space model has contributed a great deal to our understanding of the basic concepts in information retrieval [7,8].

The conventional probabilistic model offers a different approach to information retrieval. In essence, it is an adaptive model based on Bayes' decision theory. In contrast to the Boolean and vector space models, the query is not formulated directly by the user. Instead, a discriminant (decision) function representing the information request is constructed by the system through an inductive learning process (relevance feedback). Although the probabilistic model is theoretically sound, due to the problem of large

An early version of this work was presented during the New Orleans ACM SIGIR meeting, June 3–5, 1987, and appeared as "A statistical similarity measure" on pages 3–12 in *Proceedings of the Tenth Annual International ACM SIGIR Conference on Research & Development in Information Retrieval*, edited by C. T. Yu and C. J. van Rijsbergen. This final version was submitted January 27, 1988.

dimensionality one is often forced to make some rather restrictive assumptions on an nth order probability distribution of index terms (e.g. terms are assumed to be probabilistically independent in the independence model). Both the independence and other higher-order approximations [9–12] have been studied extensively. The independence model is simple to use but its validity is questionable. The tree dependence model [13] is also not very useful because it is difficult to estimate accurately the pairwise probability distributions with a small number of samples. To some extent, one can remedy this situation by enlarging the sample size. However, this makes the model impractical because the user would have to inspect a large number of documents. For these reasons, the conventional probabilistic model has gradually fallen into disfavor despite its promising beginning.

From the vantage point of autoindexing, it is possible to view each document as a probability distribution of index terms. Based on this interpretation of document representation, we propose a *probability distribution model* for information retrieval. The new approach unifies some of the key concepts in both the vector-based and existing probabilistic models. With this model, we present two plausible retrieval strategies. One is based on utility theory and the other is formulated within the framework of information theory. The new model not only enhances the retrieval effectiveness as demonstrated by our experimental results, but also provides a vigorous and practical basis for further developments in information retrieval.

2. BASIC PROBABILISTIC CONCEPTS

Before introducing the probability distribution model, we will first review briefly some of the basic concepts in probability theory.

Definition 2.1. *A system U of subsets of an arbitrary set Ω is called a σ-algebra in Ω if it has the following properties*:

(i) $\Omega \in U$,

(ii) $A \in U \Rightarrow A^c \in U$ (*A^c denotes the set complement of A*),

(iii) *for every sequence of sets $A_1, A_2, \ldots$ of U,*

$$\bigcup_{i=1}^{\infty} A_i \text{ lies in } U.$$

Let Ω be an arbitrary set and 2^Ω its power set, that is, the system of all subsets of Ω. Clearly, 2^Ω is a σ-algebra in Ω.

Definition 2.2. *Let Ω be an arbitrary set and U a σ-algebra in Ω. If a nonnegative function P, defined on the measurable space (Ω, U), satisfies the following properties*:

(i) $P(\Omega) = 1$,

(ii) *If $A_1, A_2, \ldots$ is a sequence of disjoint sets from U, then*

$$P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i),$$

the triple (Ω, U, P) is called a probability space and P a probability measure.

Each element e of Ω is called an *elementary event* that is a possible random outcome of an experiment or observation. The set of all outcomes is called the *sample space*. Each element A of U is called an *event* and $P(A)$ the probability of A.

In the present discussion, we are only interested in a special kind of probability space in which Ω is a finite set and the σ-algebra is the power set 2^Ω of Ω. In this case, for any elementary event $e \in \Omega$, the unit-set $\{e\}$ belongs to 2^Ω. Therefore, the probability $P(A)$

of any event $A \in 2^{\Omega}$ can be computed from the probabilities of the elementary events, namely,

$$P(A) = \sum_{e \in A} P(e), \tag{2.1}$$

where for notational convenience, $P(\{e\})$ is written as $P(e)$.

Definition 2.3. *Given a probability space* $(\Omega, 2^{\Omega}, P)$, *the probability measure P, defined on the measurable space* $(\Omega, 2^{\Omega})$, *is called a simple probability measure on* Ω *if*

$$P(A) = 1 \textit{ for some finite subset } A \textit{ of } \Omega.$$

Definition 2.4. *If P and Q are simple probability measures on* Ω *and* $\alpha \in [0,1]$, *the function* $\alpha P + (1-\alpha)Q$, *that assigns the value* $\alpha P(A) + (1-\alpha)Q(A)$ *to each* $A \in 2^{\Omega}$, *is called the composite distribution of P and Q.*

It can be seen that $\alpha P + (1-\alpha)Q$ is also a simple probability measure. Thus, all simple probability measures on Ω form a *convex set.*

Definition 2.5. *Suppose P is a simple probability measure on* Ω *and f a real-valued function on* Ω, *the expected value of f with respect to P is defined by*

$$E(f,P) = \sum_{e \in \Omega} f(e)P(e). \tag{2.2}$$

3. THE PROBABILITY DISTRIBUTION MODEL

3.1 *Overview*

A collection of m documents can be characterized by a set of n index terms and an $m \times n$ *document-by-term* matrix ***TF***. The element TF_{ij} of this matrix denotes the occurrence frequencies of term j in the ith document. It should, perhaps, be emphasized here that in information retrieval the problem for searching a suitable representation of documents, which truly reflects the content of each document, has not been satisfactorily resolved. Nevertheless, we assume in this article that the semantic content of each document can be approximately described by a set of index terms or keywords and the corresponding occurrence frequencies.

In the Boolean retrieval model, only partial information provided by the matrix ***TF*** is used. When evaluating the retrieval status value for any document, say d_{α}, the value of the Boolean expression t_i is defined to be *false* if $TF_{\alpha i} = 0$, and *true* whenever $TF_{\alpha i} > 0$ without being concerned with the exact value of $TF_{\alpha i}$. In the vector space model, the αth row of the matrix ***TF*** is interpreted as a vector representation of document d_{α}. That is, the element $TF_{\alpha i}$ is the ith component of the document vector $\mathbf{d}_{\alpha}$. From this point of view, the main difference between the Boolean and vector space models lies in the different interpretation and usage of the matrix ***TF***. In this article, we propose a *probabilistic* interpretation of the elements of ***TF***, which is consistent with the statistical nature of autoindexing. Each row of the matrix ***TF*** is viewed as the term probability distribution of a particular document. In other words, in the proposed probability distribution model each document is characterized by a simple probability measure on the set of index terms.

Let $T = \{t_1, t_2, \ldots, t_n\}$ be the set of terms used to index a collection of documents. One can use the relative occurrence frequencies to approximate the probability distribution. Given the matrix ***TF***, the probability $P_{d_{\alpha}}(t_i)$ of term t_i occurring in document d_{α} can be defined as follows:

$$P_{d_\alpha}(t_i) = \frac{TF_{\alpha i}}{\sum_{k=1}^{n} TF_{\alpha k}}. \tag{3.1}$$

Using the standard notation of conditional probability, $P_{d_\alpha}(t_i)$ can be written as

$$P_{d_\alpha}(t_i) = P(t_i|d_\alpha). \tag{3.2}$$

Because the power set 2^T is a σ-algebra in T, for each document d_α in a collection we can construct a probability space $(T,2^T,P_{d_\alpha})$ from the probability measure P_{d_α} defined by eqn (3.1). That is, each document is now characterized by its term distribution. Note that P_{d_α} is a simple probability measure as n is a finite natural number. Thus, it follows from eqn (2.1) that for any subset of terms $A \in 2^T$, $P_{d_\alpha}(A)$ can be computed from the formula

$$P_{d_\alpha}(A) = \sum_{t_i \in A} P_{d_\alpha}(t_i). \tag{3.3}$$

Example 3.1. Consider a 3 × 4 document-by-term matrix ***TF***:

	t_1	t_2	t_3
d_1	2	0	1
d_2	1	0	0
d_3	2	1	0

Based on eqn (3.1), each of the above documents can be represented by its term distribution as shown in the following table:

	t_1	t_2	t_3
d_1	2/3	0	1/3
d_2	1	0	0
d_3	2/3	1/3	0

Suppose $P_{d_\alpha}(t_i) > P_{d_\alpha}(t_j)$, that is, term t_i appears more often in document d_α than term t_j. One interpretation for this is that document d_α contains more information about the *concept* represented by term t_i than that of term t_j. In response to a one-term query t_i, for instance, a simple retrieval strategy is to rank the documents in decreasing order of the probabilities $P_{d_\alpha}(t_i)$, $\alpha = 1,2,\ldots,m$. However, in practical applications, a query usually consists of a set of unweighted or weighted terms. One must therefore adopt a more elaborate retrieval strategy for complex queries. Within the framework of the probability distribution model, we suggest in the following subsections two possible retrieval strategies. The first is based on a linear discriminant function that can be justified by utility theory under certain assumptions. The second method is based on a statistical similarity measure obtained from information theory.

3.2. Retrieval strategy based on utility theory

When each document d_α in a collection is characterized by a simple probability measure P_{d_α} on the set of index terms, a retrieval strategy can be formulated with the *expected-utility* model [14].

Presented with a set of documents, it is reasonable to assume that a user would prefer some documents to others according to his or her information requirements. This *user* *preference* can be described by a binary relation $\lessdot$ on Γ_s (which is the set of all simple probability measures on the set of index terms T) as follows: for $P_{d_\alpha}, P_{d_\beta} \in \Gamma_s$,

$$P_{d_\alpha} \lessdot P_{d_\beta} \Leftrightarrow \text{document } d_\beta \text{ is preferred over document } d_\alpha. \tag{3.4}$$

The concept of user preference is closely related to the more familiar concepts of *relevance* and *nonrelevance*. By definition, relevant documents are preferred by a user over nonrelevant ones. Thus, for any pair of relevant and nonrelevant documents d_β, d_α, the relationship $P_{d_\alpha} \lessdot P_{d_\beta}$ always holds.

Consider a document d_α represented by the term probability distribution P_{d_α}. If d_α is presented to a user, it is plausible that the probability for the user to obtain relevant information about a specific concept represented by term t_i is also $P_{d_\alpha}(t_i)$. Let $u(t_i)$ denote the usefulness of term t_i with respect to a particular preference relation $\lessdot$. The expected value $E(u, P_{d_\alpha})$, computed according to eqn (2.2), can be considered as a measure of the usefulness of document d_α. The above interpretation can be formally stated by the following theorem (see [14] for detailed proof).

THEOREM 3.1. *Let Γ_s be the set of all simple probability measures on T and let $\lessdot$ be a binary relation on Γ_s, representing the user preference. Then there is a real-valued function $u(t)$ on T, $u{:}T \rightarrow \mathbf{R}$, which satisfies*

$$P \lessdot Q \Leftrightarrow E(u,P) < E(u,Q), \qquad \textit{for all } P,Q \in \Gamma_s,$$

if and only if, for all $P,Q,R \in \Gamma_s$,

(i) *the relation $\lessdot$ on Γ_s is a weak order (i.e. $\lessdot$ is asymmetric and negatively transitive),*
(ii) $(P \lessdot Q,\ 0 < \alpha < 1) \Rightarrow \alpha P + (1-\alpha)R \lessdot \alpha Q + (1-\alpha)R$,
(iii) $(P \lessdot Q,\ Q \lessdot R) \Rightarrow \alpha P + (1-\alpha)R \lessdot Q$, *and* $Q \lessdot \beta P + (1-\beta)R$ *for some* $\alpha,\beta \in (0,1)$.

Moreover, $u(t)$ is uniquely defined up to a positive linear transformation. That is, $v(t)$ is also a real-valued function on T satisfying $P \lessdot Q \Leftrightarrow E(v,P) < E(v,Q)$, for all $P,Q \in \Gamma_s$, if and only if there are real numbers $a > 0$ and b such that

$$v(t) = au(t) + b \qquad \textit{for all } t \in T. \tag{3.5}$$

The function $u(t)$ is called a *utility function* defined on T and $u(t_i)$ is referred to as the *utility* of index term t_i. The *expected utility* of document d_α can now be specified by the following linear discriminant function:

$$E(u, P_{d_\alpha}) = \sum_{i=1}^{n} u(t_i) P_{d_\alpha}(t_i). \tag{3.6}$$

The implication of Theorem 3.1 is that the user preference can be described by the expected utilities of the documents under certain conditions. We can therefore rank the documents according to the values, $E(u, P_{d_\alpha})$, $\alpha = 1,2,\ldots,m$. However, there is no guarantee that the conditions in Theorem 3.1 are necessarily satisfied by an arbitrary user preference. In practice, one may not be able to vigorously justify the choice of a linear discriminant function defined by eqn (3.6) to reflect the user preference. Nevertheless, the utility theory does provide valuable insight into the retrieval process and forms a sound theoretical basis for developing an effective retrieval strategy.

So far, we have obtained the *form* of the discriminant function by making certain assumptions on the user preference relation. In general, the utility function that depends on the user preference is not known *a priori*. To specify the preference relation exactly, the user may have to read every document in the collection. Obviously, this is not a practical approach. It is therefore necessary to find a way to estimate the utility function. One may use the relevance feedback procedure to estimate the utility $u(t)$ of the individual index term as in the adaptive linear retrieval model [6]. However, it is rather difficult in practice to have an accurate estimation of $u(t)$ from a small number of samples, except perhaps in the probabilistic independence model. A simpler method is to assume that the query is input by the user as in most nonadaptive retrieval methods.

In what follows we will first show how the utility function can be estimated from an input query before discussing the relevance feedback procedure.

Estimation of the utility function from input query. Formally, a query can be considered as a real-valued function $q:T \rightarrow \boldsymbol{R}$, where T is the set of index terms. The value $q(t_i)$ specified by the user indicates the relative importance or weight of the term $t_i \in T$. From the user's point of view, terms with higher weights are regarded to contain more useful information than those with lower weights. Thus, $q(t_i)$ can be used to estimate the utility $u(t_i)$ of t_i. By approximating the utility $u(t)$ by $q(t)$, i.e. $u(t) \cong q(t)$, the expected utility of d_α can be expressed as

$$E(u,P_{d_\alpha}) = \sum_{i=1}^{n} u(t_i)P_{d_\alpha}(t_i) \cong \sum_{i=1}^{n} q(t_i)P_{d_\alpha}(t_i). \tag{3.7}$$

In general, given an input query, the documents can therefore be ranked according to the values $E(u,P_{d_\alpha})$, $\alpha = 1,2,\ldots,m$.

When the query function is restricted to take value 0 or 1, that is, $q:T \rightarrow \{0,1\}$, q is called a *binary query*. An index term t_i is considered to be useful to the user whenever $q(t_i) = 1$, and all such terms are assumed to be equally important. Such a query can be represented by the set of index terms, $A = \{t_i | q(t_i) = 1\}$. Substituting the values of the function $q(t)$ into eqn (3.7), one obtains

$$\begin{aligned} E(u,P_{d_\alpha}) &\cong \sum_{i=1}^{n} q(t_i)P_{d_\alpha}(t_i) \\ &= \sum_{t_i \in A} 1 \times P_{d_\alpha}(t_i) + \sum_{t_i \in T-A} 0 \times P_{d_\alpha}(t_i) \\ &= \sum_{t_i \in A} P_{d_\alpha}(t_i) = P_{d_\alpha}(A). \end{aligned} \tag{3.8}$$

Because the query is now characterized by the event A, in this special case we can rank the documents with respect to the probabilities $P_{d_\alpha}(A)$, $\alpha = 1,2,\ldots,m$.

EXAMPLE 3.2. Consider the document collection in Example 3.1. Suppose a query function is defined by $q(t_1) = 2$, $q(t_2) = 0$, and $q(t_3) = 1$. By the approximation, $u(t) \cong q(t)$, the expected utilities of the individual documents can be computed from eqn (3.7) as follows:

$$\begin{aligned} E(u,P_{d_1}) &= \sum_{i=1}^{3} u(t_i)P_{d_1}(t_i) \cong \sum_{i=1}^{3} q(t_i)P_{d_1}(t_i) \\ &= \left(2 \times \frac{2}{3}\right) + (0 \times 0) + \left(1 \times \frac{1}{3}\right) = \frac{5}{3}, \\ E(u,P_{d_2}) &= (2 \times 1) + (0 \times 0) + (1 \times 0) = 2, \\ E(u,P_{d_3}) &= \left(2 \times \frac{2}{3}\right) + \left(0 \times \frac{1}{3}\right) + (1 \times 0) = \frac{4}{3}. \end{aligned}$$

The ranked order of the documents is: d_2, d_1, d_3.

Estimation of the utility function from relevance feedback. Consider two documents d_α and d_β in a *training set*. If $P_{d_\alpha} \lessdot P_{d_\beta}$, it follows from Theorem 3.1 and eqn (3.6) that

$$E(u,P_{d_\alpha}) < E(u,P_{d_\beta}) \equiv \sum_{i=1}^{n} [P_{d_\beta}(t_i) - P_{d_\alpha}(t_i)]u(t_i) > 0. \tag{3.9}$$

Thus, in principle the utility function $u(t)$ in eqn (3.9) can be determined by solving a system of homogeneous linear inequalities [6]. However, due to the large number of index terms, there are three related problems inherent in such an approach. First, it requires the use of a large training set of documents that may be difficult to obtain in practice. Second, the computational complexity can easily become unmanageable. Third, they may not have a solution.

Here we suggest an alternative method to approximate the utility function $u(t)$ from a relevance feedback process. In the binary probablistic independence model, it is assumed that each document is described by a binary vector $(x_1,x_2,\ldots,x_n)$, where x_i is equal to 0 or 1 depending on the absence or presence of the index term t_i. By the Bayes' decision rule, the linear discriminant function can be expressed as:

$$g(X) = \sum_{i=1}^{n} x_i w_i = \sum_{i=1}^{n} x_i \log \frac{r_i(1 - s_i)}{(1 - r_i)s_i} + C, \tag{3.10}$$

where $r_i = P(x_i = 1|\mathit{relevant})$, $s_i = P(x_i = 1|\mathit{nonrelevant})$, and C is a constant for a given query. The value $w_i = \log[r_i(1 - s_i)/(1 - r_i)s_i]$ indicates the importance of term t_i for distinguishing relevant and nonrelevant documents.

If we assume $u(t_i) \equiv w_i$ in our probability distribution model, the expected utility of document d_i can be written as

$$E(u,P_{d_\alpha}) \cong \sum_{i=1}^{n} P_{d_\alpha}(t_i) w_i = \sum_{i=1}^{n} P_{d_\alpha}(t_i) \log \frac{r_i(1 - s_i)}{(1 - r_i)s_i}. \tag{3.11}$$

The expression defined by eqn (3.11) provides a *natural* way to incorporate term significance weights in the conventional probabilistic independence model. It is interesting to note that eqn (3.11) is in fact similar to that suggested by Croft (see eqn (9) in [15]).

3.3 *Retrieval strategy based on information theory*

In the previous subsection, we discussed within the framework of the probability distribution model how the retrieval process can be formulated according to utility theory. The main advantage of adopting utility theory to the design of retrieval strategies is that it leads to a linear discriminant function. However, the requirement that the user preference relation satisfies the conditions stated in Theorem 3.1 may be too stringent. For this reason, we propose an alternative retrieval strategy based on information theory [16]. It is assumed here that the information request is specified explicitly by the user in terms of an input query.

We have already mentioned that a query q in the utility model can be viewed as a real-valued function defined on the index term set T, namely, $q{:}T \rightarrow \boldsymbol{R}$. In this context, the user uses the value $q(t_i)$ to express the importance or usefulness of term t_i in the request for information. On the other hand, in the probability distribution model we may also view a query q as a simple probability measure P_q on the index term set T. It should be noted that such a characterization of the input query is consistent with the representation of documents as probability distributions. Thus, with respect to a query q, the probability $P_q(t_i)$ in the probability space $(T,2^T,P_q)$ provided by the user can be interpreted as the estimated probability of t_i appearing in the relevant documents. When the query is expressed in the form of natural language text, P_q can be obtained through an autoindexing process.

With this interpretation of a user query, the potential usefulness or relevance of a document can be determined by comparing its term probability distribution with that of the query. Before we introduce a statistical similarity function that measures the closeness between two probability distributions, we will first comment on the *divergence* measure originally suggested by Kullback and Leibler [17].

Consider two discrete probability distributions $P = (p_1, p_2, \ldots, p_n)$ and $Q = (q_1, q_2, \ldots, q_n)$. Suppose P is absolutely continuous with respect to Q (i.e. $p_i \rightarrow 0$ if $q_i \rightarrow 0$). The divergence $I(P, Q)$ is defined by

$$I(P,Q) = \sum_{i=1}^{n} p_i \log \frac{p_i}{q_i}. \tag{3.12}$$

Note that $I(P,Q) \geq 0$, and $I(P,Q) = 0$ if and only if P and Q are *identical* (i.e. $p_i = q_i$, $i = 1,2,\ldots,n$). From the information theory point of view, the divergence $I(P,Q)$ can be interpreted as the *difference* of the information contained in P and that contained in Q about P. If distribution Q is considered as an approximation of distribution P, the divergence $I(P,Q)$ has been used as a criterion for developing a procedure of approximating an nth order distribution by a distribution of tree dependence [13].

Let P_q and P_{d_α} be the probability distributions representing query q and document d_α. In the context of information retrieval, one would think that $I(P_q, P_{d_\alpha})$ defined in eqn (3.12) could be used to measure the *dissimilarity* between each document and the query. Unfortunately, one cannot use the divergence directly for such a purpose because P_q is not necessarily absolutely continuous with respect to P_{d_α} in a practical situation. Thus, an alternative method is needed. Based on Shannon's entropy function, we now introduce a statistical measure of similarity between two arbitrary probability distributions.

Definition 3.1. *Given a discrete probability distribution,*

$$p_i \geq 0 \;\; (i = 1,2,\ldots,n), \qquad \sum_{i=1}^{n} p_i = 1,$$

the entropy function is defined by

$$H(P) = H(p_1, p_2, \ldots, p_n) = -\sum_{i=1}^{n} p_i \log p_i. \tag{3.13}$$

For convenience, base 2 is used in the above logarithmic function. It can be easily verified that $H(P)$ has the following properties:

(i) $H(P) \geq 0$,
(ii) $H(P) = \log n$, *if* $p_1 = p_2 = , \ldots, = p_n = \frac{1}{n}$,
(iii) $H(P) = 0$, *if* $p_{i_0} = 1$ *and* $p_i = 0 \;\; (1 \leq i \leq n;\; i \neq i_0)$.

Definition 3.2. *Given two probability distributions* P, Q, *and* $\lambda_1, \lambda_2 \in [0,1]$, $\lambda_1 + \lambda_2 = 1$, *the increase of entropy for the composite distribution* $\lambda_1 P + \lambda_2 Q$ *is defined by*

$$\beta(P,Q{:}\lambda_1,\lambda_2) = H(\lambda_1 P + \lambda_2 Q) - [\lambda_1 H(P) + \lambda_2 H(Q)]. \tag{3.14}$$

Theorem 3.2. *The function* β *defined by eqn (3.14) is bounded, i.e.*

$$0 \leq \beta(P,Q{:}\lambda_1,\lambda_2) \leq 1. \tag{3.15}$$

Proof. If either λ_1 or λ_2 is equal to 0, obviously, $\beta = 0$. Suppose $\lambda_1 \neq 0$ and $\lambda_2 \neq 0$. Since $-p \log p$ is a concave function, the Jensen inequality immediately implies that the lower bound $0 \leq \beta$ holds. The upper bound $\beta \leq 1$ can be proved as follows:

$$\begin{aligned}
0 \le \beta &= H(\lambda_1 P + \lambda_2 Q) - [\lambda_1 H(P) + \lambda_2 H(Q)] \\
&= -\sum_{i=1}^{n} (\lambda_1 p_i + \lambda_2 q_i) \log (\lambda_1 p_i + \lambda_2 q_i) \\
&\quad +\lambda_1 \sum_{i=1}^{n} p_i \log p_i + \lambda_2 \sum_{i=1}^{n} q_i \log q_i \\
&= \lambda_1 \sum_{i=1}^{n} p_i \log \frac{p_i}{\lambda_1 p_i + \lambda_2 q_i} + \lambda_2 \sum_{i=1}^{n} q_i \log \frac{q_i}{\lambda_1 p_i + \lambda_2 q_i} \\
&= \lambda_1 \sum_{i=1}^{n} p_i \log \frac{\lambda_1 p_i}{\lambda_1 p_i + \lambda_2 q_i} + \lambda_2 \sum_{i=1}^{n} q_i \log \frac{\lambda_2 q_i}{\lambda_1 p_i + \lambda_2 q_i} - [\lambda_1 \log \lambda_1 + \lambda_2 \log \lambda_2].
\end{aligned}$$

Because the maximum value of $-[\lambda_1 \log \lambda_1 + \lambda_2 \log \lambda_2]$ is 1, and

$$p_i \log \frac{\lambda_1 p_i}{\lambda_1 p_i + \lambda_2 q_i} \le 0, \qquad q_i \log \frac{\lambda_2 q_i}{\lambda_1 p_i + \lambda_2 q_i} \le 0,$$

it follows that $\beta \le 1$. □

Note that $\beta(P,Q:\lambda_1,\lambda_2) = 0$ when the two probability distributions P and Q are *identical*, and the value of $\beta(P,Q:\lambda_1,\lambda_2)$ becomes maximum when P and Q are *totally different*. Thus, the entropy increase β provides a plausible measure for the *difference* between two probability distributions. (Based on an intuitive argument, β has been used to construct the distance between two random graphs in pattern recognition [18].) We show in the following theorem that there exists a close relationship between the divergence defined by eqn (3.12) and the entropy increase $\beta(P,Q:\lambda_1,\lambda_2)$.

THEOREM 3.3. $\beta(P,Q:\lambda_1,\lambda_2) = \lambda_1 I(P,\lambda_1 P + \lambda_2 Q) + \lambda_2 I(Q,\lambda_1 P + \lambda_2 Q)$. (3.16)

Proof.

$$\begin{aligned}
\beta(P,Q:\lambda_1,\lambda_2) &= H(\lambda_1 P + \lambda_2 Q) - [\lambda_1 H(P) + \lambda_2 H(Q)] \\
&= -\sum_{i=1}^{n} (\lambda_1 p_i + \lambda_2 q_i) \log (\lambda_1 p_i + \lambda_2 q_i) \\
&\quad + \lambda_1 \sum_{i=1}^{n} p_i \log p_i + \lambda_2 \sum_{i=1}^{n} q_i \log q_i \\
&= \lambda_1 \sum_{i=1}^{n} p_i \log \frac{p_i}{\lambda_1 p_i + \lambda_2 q_i} + \lambda_2 \sum_{i=1}^{n} q_i \log \frac{q_i}{\lambda_1 p_i + \lambda_2 q_i} \\
&= \lambda_1 I(P,\lambda_1 P + \lambda_2 Q) + \lambda_2 I(Q,\lambda_1 P + \lambda_2 Q).
\end{aligned}$$ □

From the above theorem, β can be viewed as the *expected divergence*. In contrast to the divergence $I(P,Q)$, both $I(P,\lambda_1 P + \lambda_2 Q)$ and $I(Q,\lambda_1 P + \lambda_2 Q)$ are well-behaved functions because P and Q are absolutely continuous with respect to $\lambda_1 P + \lambda_2 Q$. Because $I(P,\lambda_1 P + \lambda_2 Q)$ is a measure of the difference between the composite distribution $\lambda_1 P + \lambda_2 Q$ and its component P [and likewise $I(Q,\lambda_1 P + \lambda_2 Q)$ indicates the degree of difference between $\lambda_1 P + \lambda_2 Q$ and Q], the function $\beta(P,Q:\lambda_1,\lambda_2)$ can be interpreted naturally as an *indirect* measure of difference (dissimilarity measure) between the two distributions P and Q.

It should be pointed out here that the dissimilarity measure β defined by eqn (3.14) is in fact the information radius [19]. A more generalized measure based on a quadratic entropy function is given by Rao [20,21]. It is interesting to note that β can be regarded as the mutual information [12,16] between the distributions λ and $\lambda_1 P + \lambda_2 Q$ if we consider λ as the probability distribution of λ_1, λ_2.

In general, $\beta(P,Q:\lambda_1,\lambda_2)$ is not a symmetric function with respect to P and Q or with respect to λ_1 and λ_2. In information retrieval we are searching for a symmetric dissimilarity measure between P_q and P_d because we have no particular reason to emphasize either document or query. A symmetric measure can be defined as follows.

By choosing $\lambda_1 = \lambda_2 = \frac{1}{2}$, from eqn (3.14) we obtain

$$\beta\left(P,Q:\frac{1}{2},\frac{1}{2}\right) = H\left(\frac{1}{2}P + \frac{1}{2}Q\right) - \frac{1}{2}\left[H(P) + H(Q)\right]. \tag{3.17}$$

Note that in this special case, β is a *symmetric* function with respect to P and Q and the composite distribution $1/2(\mathrm{P} + \mathrm{Q})$ can be regarded as the *mean* distribution. For any probability distributions P and Q, the function $\beta(P,Q:1/2,1/2)$ satisfies the following properties:

$$\text{(i) } \beta\left(P,Q:\frac{1}{2},\frac{1}{2}\right) \geq 0,$$

$$\text{(ii) } \beta\left(P,P:\frac{1}{2},\frac{1}{2}\right) = 0,$$

$$\text{(iii) } \beta\left(P,Q:\frac{1}{2},\frac{1}{2}\right) = \beta\left(Q,P:\frac{1}{2},\frac{1}{2}\right).$$

From Theorem 3.2, we also know that β is bounded (i.e. $0 \leq \beta \leq 1$). Therefore, a similarity measure between two probability distributions P and Q can be defined as

$$\mathrm{SIM}(P,Q) = 1 - \beta\left(P,Q:\frac{1}{2},\frac{1}{2}\right). \tag{3.18}$$

Example 3.3. This example illustrates the basic concepts of the statistical similarity measure introduced in this section. Figure 1(a) shows the similarity between two identical probability distributions, Figure 1(b) shows the similarity between two very different probability distributions, and Figure 1(c) shows the similarity of two arbitrary probability distributions.

In the proposed probability distribution model, the function defined by eqn (3.18) can

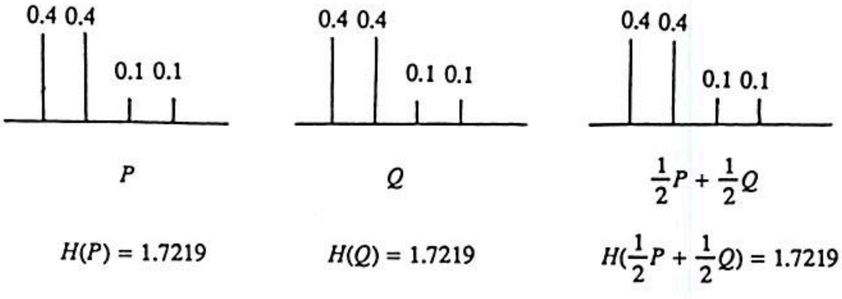

Fig. 1(a). $\beta = 0$, $SIM(P,Q) = 1$

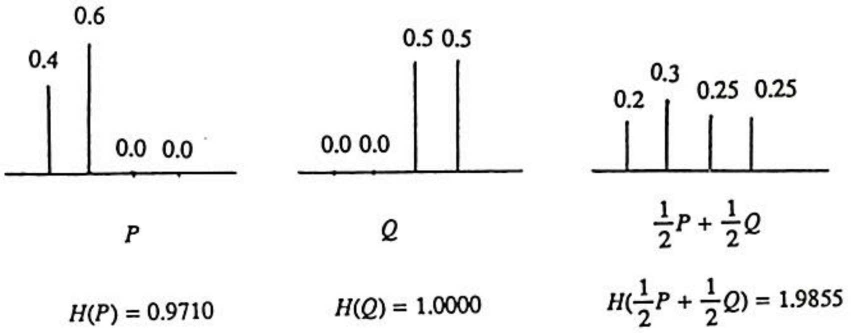

Fig. 1(b). $\beta = 1$, $SIM(P,Q) = 0$

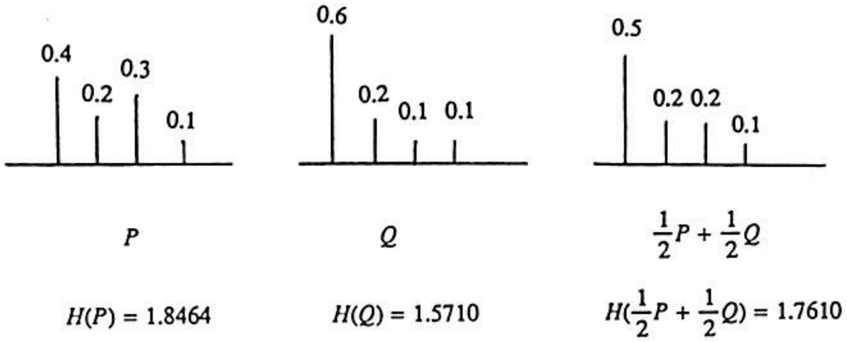

Fig. 1(c). $\beta = 0.0523$, $SIM(P,Q) = 0.9477$

be adopted as a similarity measure between a document and a query. Given document d_α and query q, the similarity function is defined by

$$\mathrm{SIM}(P_{d_\alpha},P_q) = 1 - \beta\left(P_{d_\alpha},P_q: \frac{1}{2}, \frac{1}{2}\right). \tag{3.19}$$

Example 3.4. Let a query q be expressed as a probability distribution:

$$P_q(t_1) = \frac{2}{3}, \qquad P_q(t_2) = 0, \qquad P_q(t_3) = \frac{1}{3}.$$

From the document collection given in Example 3.1, the similarity between the two probability distributions P_q and P_{d_2} can be calculated in the following steps:

(i) $\frac{1}{2} P_{d_2} + \frac{1}{2} P_q = \frac{1}{2}(1,0,0) + \frac{1}{2}\left(\frac{2}{3},0,\frac{1}{3}\right) = \left(\frac{5}{6},0,\frac{1}{6}\right),$

(ii) $H(P_{d_2}) = -(1\log 1 + 0\log 0 + 0\log 0) = 0.0000,$

$$H(P_q) = -\left(\frac{2}{3}\log\frac{2}{3} + 0\log 0 + \frac{1}{3}\log\frac{1}{3}\right) = 0.9183,$$

$$H\left(\frac{1}{2}P_{d_2} + \frac{1}{2}P_q\right) = -\left(\frac{5}{6}\log\frac{5}{6} + 0\log 0 + \frac{1}{6}\log\frac{1}{6}\right) = 0.6500,$$

$$\text{(iii) } \mathrm{SIM}(P_{d_2},P_q) = 1 - \beta\left(P_{d_2},P_q\colon \frac{1}{2},\frac{1}{2}\right)$$

$$= 1 - \left[H\left(\frac{1}{2}P_{d_2} + \frac{1}{2}P_q\right) - \frac{1}{2}\left(H(P_{d_2}) + H(P_q)\right)\right]$$

$$= 1 - [0.6500 - 0.5 \times (0.0000 + 0.9183)] = 0.8092.$$

The other similarity values can be calculated in the same way. Based on the following similarity values:

$$\mathrm{SIM}(P_{d_1},P_q) = 1.0000,$$

$$\mathrm{SIM}(P_{d_2},P_q) = 0.8092,$$

$$\mathrm{SIM}(P_{d_3},P_q) = 0.6667,$$

the ranked order of the documents is: d_1, d_2, d_3.

The similarity measure defined in eqn (3.19) is applied to a number of test document collections and the results are presented in the next section.

4. EXPERIMENTAL RESULTS

Four document collections were used in our experiments: ADINUL, CRN4NUL, MEDNUL, and MEDLARS. The characteristics of these collections are listed below:

- ADINUL is a collection of 82 documents in library science. It consists of the full text of papers presented at the 1963 meeting of the American Documentation Institute. There are 35 queries in the query collection.
- CRN4NUL has 424 abstracts on aerodynamics, which were used by the Cranfield Project. The corresponding query collection involves 155 queries.
- MEDNUL is a collection of 450 documents and 30 queries. The documents are in the area of biomedicine.
- MEDLARS is a collection with 1,033 documents, also in biomedicine, and has 30 queries associated with it.

The indexing of the first three collections is done automatically in the SMART system [7] using the word-stem method. The last two collections are subsets of documents prepared by the National Library of Medicine. The query collections include, for evaluation purposes, information as to which documents are relevant to each query.

The standard recall and precision measures are used for comparing the performance of different models. Recall is defined as the proportion of relevant documents retrieved and precision is the proportion of the retrieved documents acutally relevant. The overall performance of each model is determined by computing the average precision over all the queries for recall values 0.1, 0.2, . . . , and 1.0. The algorithm for averaging is consistent with that implemented in the SMART system.

We performed two groups of experiments and the results are presented as follows:

4.1 *Comparison of linear and nonlinear discriminant functions in the Probability Distribution Model (PDM)*

In the first group of experiments, the objective is to compare the retrieval effectiveness of the linear discriminant function [eqn (3.7)] derived from the utility theory with that of the nonlinear statistical similarity measure [eqn (3.19)]. In the former case, the weight $q(t)$ of each index term in the input query is used as an approximation of the utility $u(t)$, whereas in the latter ranking strategy based on the information theory, the query is treated as a probability distribution of index terms. The experimental results are summarized in Table 1. In all four document collections the retrieval results of the statistical approach are significantly better than those obtained from the linear discriminant function. In the MEDLARS collection we obtained an average improvement of 14% and 26% in the ADINUL collection. The improvements are more pronounced in the lower recall values for both the ADINUL and CRN4NUL collections. Our analysis here suggests that adopting a linear similarity measure in information retrieval can, perhaps, be viewed as a first order of approximation. In general, nonlinear similarity measures should perform better than linear ones.

4.2 *Comparison of the cosine similarity in VSM with the linear discriminant function in PDM*

The objective of the second group of our experiments is to compare two linear retrieval strategies. One is based on the *cosine* similarity measure in the standard vector space model (VSM) and the other is based on the linear discriminant function (the expected utility) in the probability distribution model. It can be seen from Table 2 that there is no significant difference in the performance results between the standard vector space model and the probability distribution model with the adoption of a linear discriminant function. This may be attributed to the fact that they both use a linear similarity measure.

In contrast, by comparing the results in Tables 1 and 2, one can see that by adopt-

Table 1. Comparison of linear and nonlinear retrieval strategies

ADINUL (82 docs, 35 ques)

	Precision		
Recall	Linear	Nonlinear	%
0.10	0.3262	0.4531	38.9
0.20	0.3049	0.4202	37.8
0.30	0.2825	0.3836	35.8
0.40	0.2418	0.3156	30.5
0.50	0.2365	0.3103	31.2
0.60	0.1825	0.2220	21.6
0.70	0.1437	0.1602	11.5
0.80	0.1306	0.1565	19.8
0.90	0.1070	0.1252	17.0
1.00	0.1060	0.1228	15.8
Average % improvement			26.0

CRN4NUL (424 docs, 155 ques)

	Precision		
Recall	Linear	Nonlinear	%
0.10	0.5751	0.6644	15.5
0.20	0.5013	0.6051	20.7
0.30	0.4118	0.5091	23.6
0.40	0.3225	0.3926	21.7
0.50	0.2882	0.3614	25.4
0.60	0.2368	0.2883	21.7
0.70	0.1836	0.2069	12.7
0.80	0.1510	0.1691	12.0
0.90	0.1160	0.1290	11.2
1.00	0.1125	0.1241	10.3
Average % improvement			17.5

MEDNUL (450 docs, 30 ques)

	Precision		
Recall	Linear	Nonlinear	%
0.10	0.7199	0.8134	13.0
0.20	0.6038	0.7063	17.0
0.30	0.4970	0.5809	16.9
0.40	0.4349	0.5114	17.6
0.50	0.4021	0.4645	15.5
0.60	0.3308	0.3857	16.6
0.70	0.3052	0.3583	17.4
0.80	0.2280	0.2676	17.4
0.90	0.1633	0.2029	24.2
1.00	0.1191	0.1417	19.0
Average % improvement			17.5

MEDLARS (1,033 docs, 30 ques)

	Precision		
Recall	Linear	Nonlinear	%
0.10	0.7533	0.8422	11.8
0.20	0.6409	0.7167	11.8
0.30	0.5558	0.6316	13.6
0.40	0.4758	0.5732	20.5
0.50	0.4058	0.4892	20.6
0.60	0.3602	0.4216	17.0
0.70	0.3091	0.3463	12.0
0.80	0.2625	0.2783	6.0
0.90	0.1508	0.1720	14.1
1.00	0.0710	0.0781	10.0
Average % improvement			13.7

Table 2. Cosine similarity (VSM) vs. linear discriminant function (PDM)

ADINUL (82 docs, 35 ques)

	Precision		
Recall	Cosine	Linear	%
0.10	0.3786	0.3262	−13.8
0.20	0.3434	0.3049	−11.2
0.30	0.3094	0.2825	−8.7
0.40	0.2587	0.2418	−6.5
0.50	0.2465	0.2365	−4.1
0.60	0.1887	0.1825	−3.3
0.70	0.1357	0.1437	5.9
0.80	0.1283	0.1306	1.8
0.90	0.1098	0.1070	−2.6
1.00	0.1082	0.1060	−2.0
Average % improvement			−4.5

CRN4NUL (424 docs, 155 ques)

	Precision		
Recall	Cosine	Linear	%
0.10	0.6415	0.5751	−10.4
0.20	0.5540	0.5013	−9.5
0.30	0.4514	0.4118	−8.8
0.40	0.3621	0.3225	−10.9
0.50	0.3250	0.2882	−11.3
0.60	0.2726	0.2368	−13.1
0.70	0.2059	0.1836	−10.8
0.80	0.1655	0.1510	−8.8
0.90	0.1240	0.1160	−6.5
1.00	0.1179	0.1125	−4.6
Average % improvement			−9.5

MEDNUL (450 docs, 30 ques)

	Precision		
Recall	Cosine	Linear	%
0.10	0.7510	0.7199	−4.1
0.20	0.6407	0.6038	−5.8
0.30	0.5254	0.4970	−5.4
0.40	0.4687	0.4349	−7.2
0.50	0.4332	0.4021	−7.2
0.60	0.3707	0.3308	−10.8
0.70	0.3394	0.3052	−10.1
0.80	0.2400	0.2280	−5.0
0.90	0.1796	0.1633	−9.1
1.00	0.1231	0.1191	−3.2
Average % improvement			−6.8

MEDLARS (1,033 docs, 30 ques)

	Precision		
Recall	Cosine	Linear	%
0.10	0.7824	0.7533	−3.7
0.20	0.6931	0.6409	−7.5
0.30	0.5879	0.5558	−5.5
0.40	0.5450	0.4758	−12.7
0.50	0.4409	0.4058	−8.0
0.60	0.3821	0.3602	−5.7
0.70	0.3296	0.3091	−6.2
0.80	0.2706	0.2625	−3.0
0.90	0.1547	0.1508	−2.5
1.00	0.0832	0.0710	−14.7
Average % improvement			−6.9

ing a nonlinear similarity measure the average performance of the probability distribution model is about 10% better than that of the vector space model. In particular, an average improvement of 20% is observed in the ADINUL collection.

5. CONCLUSIONS

We have presented in this article a probability distribution model for information retrieval. One of the salient features of this model is that it brings together many basic concepts introduced over the years in a variety of retrieval models.

In our model, each document is uniquely characterized by a simple probability measure on the set of index terms. Based on this new interpretation of document representation, we have suggested two possible retrieval strategies. One is a linear discriminant function based on utility theory and the other is a nonlinear statistical similiarity measure based on information theory.

In order to evaluate the retrieval effectiveness of the proposed model, several test collections of documents were used in our experiments. The experimental results indicate that there is no substantial difference in performance between the standard vector space model and the linear retrieval strategy adopted by the probability distribution model. However, it can be clearly seen that the statistical similarity measure produces consistently better results than the usual cosine similarity function.

In this article we have discussed document representation and retrieval strategies based entirely on the concepts represented by the *index terms*. In fact, it may be more appropriate to use the *atomic* concepts introduced in the generalized vector space model [2,3,22] for the development of the probability distribution model. With the atomic concepts, the formulism represented here can be easily extended to provide facilities for processing structured queries as well. The reason why we have chosen to present the new model in terms of the term concepts is primarily for the sake of clarity.

Acknowledgement—We would like to thank P. Bollmann for his valuable suggestions on the expected-utility model.

REFERENCES

1. Raghavan, V.V.; Wong, S.K.M. A critical analysis of vector space model in information retrieval. Journal of the American Society for Information Science, 37: 279–287; 1986.
2. Wong, S.K.M.; Ziarko, W.; Wong, P.C.N. Generalized vector space model in information retrieval. Proceedings of the ACM SIGIR conference on research and development in information retrieval; 1985; 18–25.
3. Wong, S.K.M.; Ziarko, W.; Raghavan, V.V.; Wong, P.C.N. On extending the vector space model for Boolean query processing. Proceedings of the ACM SIGIR conference on research and development in information retrieval; 1986; 175–185.
4. Yu, C.T. A formal construction of term classes. Journal of the ACM, 22: 17–37; 1975.
5. Raghavan, V.V.; Yu, C.T. Experiments on the determination of the relationships between terms. ACM Transactions on Database Systems, 4: 240–260; 1979.
6. Bollmann, P.; Wong, S.K.M. Adaptive linear information retrieval models. Proceedings of the ACM SIGIR conference on research and development in information retrieval; 1987.
7. Salton, G., editor. The SMART retrieval system—experiments in automatic document processing. Englewood Cliffs, NJ: Prentice-Hall; 1971.
8. Salton, G; McGill, M.H. Introduction to modern information retrieval. New York: McGraw-Hill; 1983.
9. Yu, C.T.; Salton, G. Precision weighting—an effective automatic indexing method. Journal of the ACM, 23: 76–85; 1976.
10. Yu, C.T.; Luk, W.S.; Cheung, T.Y. A statistical model for relevance feedback in information retrieval. Journal of the ACM, 23: 273–286; 1976.
11. Robertson, S.E.; Sparck Jones, K. Relevance weighting of search terms. Journal of the American Society for Information Science, 27: 129–146; 1976.
12. Van Rijsbergen, C.J. Information retrieval. London: Butterworths; 1980.
13. Chow, C.K.; Liu, C.N. Approximating discrete probability distributions with dependence trees. IEEE Transactions on Information Theory, IT-14: 462–467; 1968.
14. Fishburn, P.C. Utility theory for decision making. New York: Wiley; 1970.
15. Croft, W.B. Document representation in probabilistic models of information retrieval. Journal of the American Society for Information Science, 32: 451–457; 1981.
16. Aczel, J; Daroczy, Z. On measures of information and their characterizations. Mathematics in science and engineering, vol. 115. New York: Academic Press; 1975.
17. Kullback, S.; Leibler, R.A. On information and sufficiency. Annual Mathematical Statistics, 22: 79–86; 1951.
18. Wong, A.K.C.; You, M. Entropy and distance of random graphs with application to structural pattern recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, PAMI-7: 599–609; 1985.
19. Jardine, N.; Sibson, R. Mathematical taxonomy. New York: Wiley; 1971.
20. Rao, C.R.; Nayak, T.K. Cross entropy, dissimilarity measures, and characterizations of quadratic entropy. IEEE Transactions on Information Theory, IT-21: 589–593; 1985.
21. Rao, C.R. Diversity and dissimilarity coefficients: A unified approach. Theoretical Population Biology, 21: 24–43; 1982.
22. Wong, S.K.M.; Yao, Y.Y. A statistical similarity measure. Proceedings of the ACM SIGIR conference on research and development in information retrieval; 1987.