A Note on Inverse Document Frequency Weighting Scheme

S. K. M. Wong
Y. Y. Yao

TR 89-990
April 1989

Department of Computer Science
Cornell University
Ithaca, NY 14853-7501

A Note on Inverse Document Frequency Weighting Scheme

S.K.M. Wong and Y.Y. Yao

Abstract

Based on the Shannon information theory, a measure for term value is introduced. This study is an attempt to provide a theoretical justification for the inverse document frequency (IDF) weighting scheme. The argument presented in this paper is somewhat different from those suggested earlier. It is shown that IDF weights can be derived from the proposed approach by assuming that each index term has an even distribution within a subset of documents. A critical comment on the signal-noise ratio (S/N) weighting method is also included.

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

1. Introduction

One of the fundamental and perhaps most difficult issues in information retrieval is content analysis. There have been extensive studies on document indexing in which a set of index terms is chosen to represent the content of the individual document [8,12,14,18]. An index term fulfills two dual functions, namely, recall and precision. From the point of view of recall, a *good* term should be a descriptor which accurately reflects the content of the individual document. A set of carefully selected index terms should enhance the likelihood of retrieving relevant documents [8,11,14,18] or ranking the more preferred documents ahead of the less preferred [19,20]. As for precision, one should eliminate redundant index terms in order to prevent indiscriminate retrieval. Thus, to judge the *goodness* (value) of an index term depends on how well it performs both the recall and precision functions. Many attempts have been made to relate the value of a term to its occurrence frequency within the documents in the whole collection [12,14]. These investigations have led to many useful term weighting schemes. For example, the usefulness of a term for describing the content of a document can be estimated from the term significance weighting methods [3,12], while the effectiveness of a term for differentiating documents can be measured by the term precision weighting schemes [11,22]. One may compute term weights directly from the probabilistic model [11,18] or from the adaptive linear model [19]. Term precision weights can also be estimated by using various methods such as the inverse document frequency (IDF) [1,9,13,16], the signal-noise ratio (S/N) [5,12,14], and the term discrimination [12,14] weighting schemes. The evidence accumulated so far clearly indicates that term weighting provides an effective means to improve the performance of an information retrieval system [12-14,17].

The main drawback of the IDF weighting scheme is that it ignores the information about relative term frequencies. That is, whenever two terms occur in the same number of documents, they are assigned the same IDF weight. In contrast, the S/N weighting method attempts to take into account the effect of relative term occurrence frequencies.

Unfortunately, it was shown [12,14] that the S/N method produces inferior performance results than the IDF method.

Based on the Shannon information theory [2,6,15], we present a measure to evaluate the goodness of a term. This study is an attempt to provide a theoretical justification for the IDF weighting scheme. Our argument is somewhat different from those suggested earlier [4,21]. We show that IDF weights can be derived from the proposed approach by assuming that each index term occurs uniformly within a subset of documents. We also comment on the relationship between the S/N and the proposed methods.

2. A Measure of Evidence

Before discussing the notion of term weighting in information retrieval, we first introduce a measure to estimate the *goodness* of a piece of evidence based on the Shannon information theory [15].

Consider a probability space $(P, \Theta, 2^{\Theta})$ with P as the probability function. Let $D = \{d_1, d_2, \ldots, d_N\}$ be a set of mutually exclusive *propositions* (events) satisfying:

$$\bigcup_{i=1}^{N} d_i = \Theta \ . \tag{2.1}$$

For any *evidence* $t \subseteq \Theta$, we can define a probability distribution as follows:

$$p_i = P(d_i \mid t) \ \text{ for } i = 1, 2, \ldots, N, \quad p_i \geq 0 \ \text{ and } \ \sum_{i=1}^{N} p_i = 1 \ , \tag{2.2}$$

where $P(d_i \mid t)$ denotes the conditional probability. The corresponding entropy (Shannon's entropy) is the quantity:

$$H(P_t) = H(p_1, p_2, \ldots, p_N) = -\sum_{i=1}^{N} p_i \log p_i \ , \tag{2.3}$$

where $p_i \log p_i = 0$ if $p_i = 0$, extending $-x \log x$ to the origin by continuity.

The entropy function $H(P_t)$ satisfies the following properties:

(i) $0 \le H(P_t) \le \log N$,

(ii) $H(P_t) = \log N$, if $p_1 = p_2 = \cdots = p_n = \frac{1}{N}$,

(iii) $H(P_t) = 0$, if $p_{i_0} = 1$ and $p_i = 0$ $(1 \le i \le N,\ i \ne i_0)$.

Figure 1 depicts the entropy function $H(P_t) = p \log p + (1-p) \log (1-p)$. It can be seen that $H(P_t)$ reaches the maximum value when P_t becomes an *even* distribution (i.e. $p = 1/2$). Let us consider an example in which $D = \{d_1, d_2, d_3\}$, and

$$
\begin{aligned}
P_{t_1} &= (P(d_1 \mid t_1), P(d_2 \mid t_1), P(d_3 \mid t_1)) = (0.33, 0.33, 0.34) , \\
P_{t_2} &= (P(d_1 \mid t_2), P(d_2 \mid t_2), P(d_3 \mid t_2)) = (0.80, 0.20, 0.00) , \\
P_{t_3} &= (P(d_1 \mid t_3), P(d_2 \mid t_3), P(d_3 \mid t_3)) = (0.60, 0.00, 0.40) .
\end{aligned}
$$

are three different probability distributions defined by evidence t_1, t_2, and t_3, respectively. A natural question to ask is: which body of evidence reduces more uncertainty with regard to the propositions in D ? We attempt to answer this question by introducing a measure for evaluating evidence.

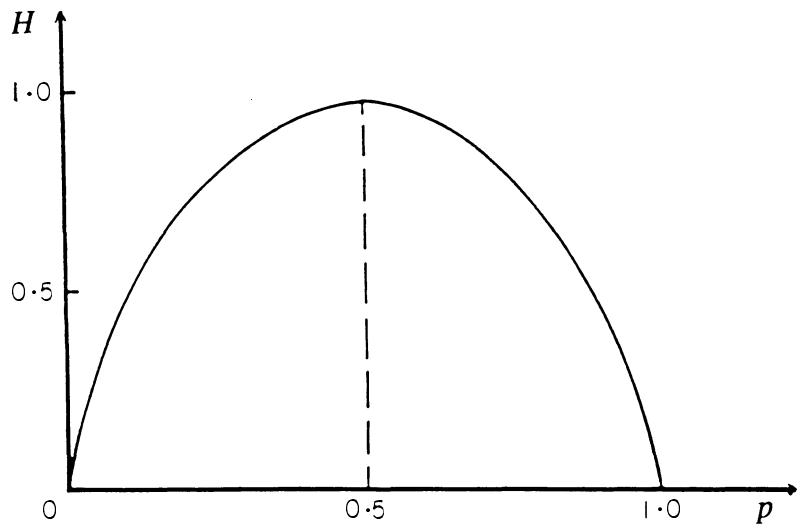

Figure 1. Entropy function $H(p,\ 1-p)$

For a given set D of propositions, the entropy $H(P_t)$ defined by eqns. (2.2) and (2.3) can be interpreted as a measure of the level of information uncertainty based on what we know about evidence t. Consider a finite set of evidence, $T = \{t_1, t_2, \ldots, t_M\}$. Suppose $H(P_{t_i}) > H(P_{t_j})$ for a pair of evidence $t_i, t_j \in T$. One may conclude that evidence t_j is *better* than evidence t_i because t_j produces a state of lower uncertainty. Since we are primarily interested in comparing the relative ability of different bodies of evidence to reduce the level of uncertainty, given evidence t we define the following measure of *uncertainty*:

$$\beta(P_t) = \frac{H(P_t)}{H_{\max}} \quad , \quad (0 \le \beta(P_t) \le 1), \tag{2.4}$$

where $H_{\max}$ is the maximum entropy among the entropies, $H(P_{t_1}), H(P_{t_2}), \ldots, H(P_{t_M})$ defined by the set T of evidence. The ratio $\beta(P_t)$ is sometimes referred to as the *relative entropy* [15], which measures how uncertain we are of the propositions in D when we are given evidence t. Alternatively, one may define given evidence t a measure of *certainty* as follows:

$$w_t = 1 - \beta(P_t) = 1 - \frac{H(P_t)}{H_{\max}} \quad , \quad (0 \le w_t \le 1) \ . \tag{2.5}$$

This function is used in subsequent discussions as a measure of the *goodness* (value) of evidence t. Applying this measure to the set of evidence $\{t_1, t_2, t_3\}$ considered in the earlier example, it can be easily verified that $w_{t_2} > w_{t_3} > w_{t_1}$.

It is perhaps worth mentioning that a measure similar to w_t is used in information theory to estimate the *redundancy* of a language or an information source [6,15]. Such a measure is also used to assess the usefulness of an *attribute* in the study of multi-attributes decision making [7,23]. In the following section, we apply the measure of evidence defined by eqn. (2.5) to term precision weighting in information retrieval and comment on other related weighting schemes.

3. Term Precision Weighting

In the context of information retrieval, an index term may be interpreted as evidence, and documents as propositions. Based on such an interpretation, a measure of the value of evidence can be used as a measure of term value.

Suppose a collection of N documents is characterized, through M index terms, by a $N \times M$ *document–by–term* matrix TF. The element TF_{ij} of this matrix denotes the occurrence frequency of term t_j in document d_i. Let $TOLTF_j$ denote the total occurrence frequency of term t_j in the whole collection, namely:

$$TOLTF_j = \sum_{i=1}^{N} TF_{ij} \ . \tag{3.1}$$

If one assumes:

$$p_{ij} = P(d_i \mid t_j) = \frac{TF_{ij}}{TOLTF_j} \geq 0 \quad \text{for } i = 1, 2, \ldots, N \ , \tag{3.2}$$

then one may define for each index term t_j a probability distribution:

$$P_{t_j} = (p_{1j}, p_{2j}, \ldots, p_{Nj}) = \left(\frac{TF_{1j}}{TOLTF_j}, \frac{TF_{2j}}{TOLTF_j}, \ldots, \frac{TF_{Nj}}{TOLTF_j}\right) \ . \tag{3.3}$$

Based on eqns. (2.5) and (3.3), the *value* of term t_j is given by:

$$w_{t_j} = 1 - \frac{H(P_{t_j})}{H_{\max}} \ , \tag{3.4}$$

where $H_{\max}$ is the maximum entropy. For a collection of N documents, according to property (iii) satisfied by an entropy function, the maximum value $H_{\max}$ is $\log N$. In this case, we obtain the following measure of term value:

$$w_{t_j} = 1 - \frac{H(P_{t_j})}{\log N} \ . \tag{3.5}$$

However, if certain restriction is imposed on term occurrence in a document collection, the maximum entropy $H_{\max}$ may differ from $\log N$. For example, in some automatic indexing methods common functional words are deleted from the index vocabulary [14,18]. This in fact implicitly imposes a condition that no index term can occur in more than *max n* documents. Under this condition, the maximum value $H_{\max}$ is $\log max\ n$. Then the new measure (3.4) can be written as:

$$w_{t_j} = 1 - \frac{H(P_{t_j})}{\log max\ n} . \tag{3.6}$$

IDF weighting scheme

The inverse document frequency (IDF) scheme is perhaps the best known method for term weighting. We wish to show that the IDF method is a special case of the proposed measure of term value.

Let n_j denote the number of documents indexed by t_j, which is the number of non-zero entries in the jth column of the *TF* matrix. (n_j is often called the document frequency of term t_j.) Sparck Jones [9,16,17] suggested that t_j may be weighted by the function:

$$w'_{t_j} = -\log \frac{n_j}{N} , \tag{3.7}$$

where N is the number of documents in the collection. This formula was implemented in some experiments [17] as

$$w'_{t_j} = -\log \frac{n_j}{max\ n} . \tag{3.8}$$

Both formulas (3.7) and (3.8) implicitly assume that the *TF* matrix is a *binary* matrix by neglecting the relative term occurrence frequency information. Let D_{t_j} denote the subset of documents indexed by t_j in the collection. The binary interpretation of the

TF matrix in effect assumes that each index term t_j is characterized by an even probability distribution in D_{t_j}, namely:

$$P_{t_j} = (p_{1j}, p_{2j}, \ldots, p_{Nj}), \qquad p_{ij} = \begin{cases} 1/n_j & d_i \in D_{t_j}, \\ 0 & \textit{otherwise}. \end{cases} \tag{3.9}$$

By substituting eqn. (3.9) into eqn. (2.3), one obtains:

$$\begin{aligned} H(P_{t_j}) &= -\sum_{i=1}^{N} p_{ij} \log p_{ij} = -\sum_{d_i \in D_{t_j}} \frac{1}{n_j} \log \frac{1}{n_j} \\ &= -n_j \left(\frac{1}{n_j} \log \frac{1}{n_j} \right) = \log n_j . \end{aligned} \tag{3.10}$$

Thus, according to eqn. (3.5), the value of term t_j is given by:

$$\begin{aligned} w_{t_j} &= 1 - \frac{\log n_j}{\log N} \\ &= \frac{1}{\log N} [\log N - \log n_j] \\ &= \frac{1}{\log N} [-\log \frac{n_j}{N}] . \end{aligned} \tag{3.11}$$

Since $1/\log N$ is a positive constant independent of any particular term, we can immediately conclude that eqn. (3.11) is equivalent to formula (3.7). If one imposes the restriction that any term in the index vocabulary cannot occur in more than *max n* documents, a weighting formula equivalent to eqn. (3.8) can be easily derived from eqn. (3.6).

The argument presented has shown that the IDF weighting scheme is indeed a special case of the proposed weighting method. It should be emphasized that using n_j (the document frequency of t_j) instead of the actual term distribution P_{t_j} to estimate the value of term t_j may be viewed as a first order approximation. This important point is clarified in the following example.

Example 1

Consider the document-by-term matrix *TF*:

	t_1	t_2	t_3	$\cdots$	t_M
d_1	1	1	2	$\cdots$	·
d_2	1	1	4	$\cdots$	·
d_3	1	0	0	$\cdots$	·
d_4	1	2	2	$\cdots$	·
d_5	0	2	0	$\cdots$	·
d_6	0	0	4	$\cdots$	·
d_7	0	0	0	$\cdots$	·
·	·	·	·	$\cdots$	·
d_N	0	0	0	$\cdots$	·

As defined by eqn. (3.3), terms t_1, t_2, and t_3 in the above collection are characterized by the probability distributions:

$$P_{t_1} = (1/4,\ 1/4,\ 1/4,\ 1/4,\ 0, \ldots, 0) ,$$
$$P_{t_2} = (1/6,\ 1/6,\ 0,\ 2/6,\ 2/6,\ 0, \ldots, 0) ,$$
$$P_{t_3} = (2/12,\ 4/12,\ 0,\ 2/12,\ 0,\ 4/12,\ 0, \ldots, 0) .$$

By using the IDF weighting scheme (3.7), the *same* value $(-\log 4/N)$ is assigned to t_1, t_2, and t_3 (i.e. $w'_{t_1} = w'_{t_2} = w'_{t_3}$). On the other hand, in our method the values of these three terms are computed from eqn. (3.5):

$$H(P_{t_1}) = -(\frac{1}{4}\log\frac{1}{4} + \frac{1}{4}\log\frac{1}{4} + \frac{1}{4}\log\frac{1}{4} + \frac{1}{4}\log\frac{1}{4}) = \log 4 ,$$
$$w_{t_1} = 1 - \frac{H(P_{t_1})}{\log N} = 1 - \frac{\log 4}{\log N} = \frac{1}{\log N}[\log N - \log 4] ,$$

$$H(P_{t_2}) = -(\frac{1}{6}\log\frac{1}{6} + \frac{1}{6}\log\frac{1}{6} + \frac{2}{6}\log\frac{2}{6} + \frac{2}{6}\log\frac{2}{6}) = \log 6 - \frac{2}{3}\log 2 ,$$
$$w_{t_2} = 1 - \frac{H(P_{t_2})}{\log N} = 1 - \frac{\log 6 - \frac{2}{3}\log 2}{\log N} = \frac{1}{\log N}[\log N - (\log 6 - \frac{2}{3}\log 2)] .$$

Obviously, P_{t_2} and P_{t_3} are characterized by the same entropy, so the same weight is

assigned to both t_2 and t_3 (i.e. $w_{t_2} = w_{t_3}$). In contrast to the IDF method, based on the new measure both t_2 and t_3 have a higher value than t_1 (i.e. $w_{t_1} < w_{t_2} = w_{t_3}$).

The proposed measure for term value is developed from the concepts of the probability and information theories. Our argument is not based on the results of the probabilistic independence model. It is perhaps useful to look at some earlier theoretical arguments for the IDF weighting scheme.

In the binary probabilistic independence model [11,18], with respect to a given query, term t_j is assigned a weight:

$$w''_{t_j} = \log \frac{p_i(1-q_j)}{q_i(1-p_j)} , \tag{3.12}$$

where p_j is the probability that a relevant document contains term t_j, and q_j is the probability that a non-relevant document contains term t_j. Croft and Harper [4] derived a formula (see eqn. (3.14) below) similar to eqn. (3.7) as follows. They argued that when the relevance information is not available, there is not much one can say about p_j, and it would therefore seem appropriate to assume that p_j is constant. On the other hand, because the number of relevant documents is usually small compared to the total collection size, q_j can be approximated by the relative occurrence frequency, n_j/N. Thus, it follows:

$$w''_{t_j} \approx \log \frac{p_j}{1-p_j} + \log \frac{N-n}{N} . \tag{3.13}$$

Furthermore, if one assumes that $p_j = 0.5$, eqn. (3.13) immediately reduces to

$$w''_{t_j} \approx \log \frac{N-n}{N} . \tag{3.14}$$

An interesting point is that despite these crude assumptions, eqn. (3.14) is *not* exactly equivalent to formula (3.7) as originally suggested by Sparck Jones [16,17].

Robertson [10] pointed out that some of the arguments for the IDF weighting scheme within the framework of the probabilistic independence model are not only inconclusive but also confusing. For example, consider two other weighting formulas [11]:

$$w_{t_j}^1 = \log \frac{\frac{r_j}{R}}{\frac{n_j}{N}} , \tag{3.15}$$

$$w_{t_j}^3 = \log \frac{\frac{r_j}{R - r_j}}{\frac{n_j}{N - n_j}} , \tag{3.16}$$

where r_j is the number of relevant documents indexed by t_j, and R is the number of relevant documents. If one assumes that $p_j = n_j/R = 1$, eqn. (3.15) would reduce to the formula used by Sparck Jones. On the other hand, by assuming that $p_j = 0.5$, eqn. (3.16) would reduce to eqn. (3.14) as suggested by Croft and Harper. It is interesting to note that Wu and Salton [21] presented another explanation for the IDF formula by using a different set of assumptions. Based on these findings, an inevitable question arises: which of these weighting functions and assumptions should be combined to properly explain the IDF weighting scheme ?

S/N weighting scheme

Here we compare the signal-noise ratio (S/N) scheme [12,14] with the proposed method. Given the probability distribution P_{t_j}, the value of t_j in the S/N scheme is measured by:

$$S_{t_j} = \log TOLTF_j - H(P_{t_j}) , \tag{3.17}$$

where $TOLTF_j$ is the total occurrence frequency of t_j in the whole collection. A detailed

discussion of this method can be found in [12,14].

According to eqn. (3.5), the term value of t_j is given by:

$$w_t = 1 - \frac{H(P_{t_j})}{\log N} = \frac{1}{\log N} [\log N - H(P_{t_j})] \ . \tag{3.18}$$

Since we are primarily interested in *relative* term weights, the constant factor $1/\log N$ in eqn. (3.18) can be ignored. Therefore, the following equivalent weighting formula may be used:

$$w_{t_j} = \log N - H(P_{t_j}) \ . \tag{3.19}$$

The only difference between eqns. (3.17) and (3.19) is that $\log TOLTF_j$ is used in the former while $\log N$ is used in the latter. An example is given below to show some of the drawbacks in the S/N method.

Example 2

Consider the TF matrix given in Example 1. From eqn. (3.19), we obtain:

$$w_{t_1} = \log N - \log 4 \ ,$$

$$w_{t_2} = w_{t_3} = \log N - (\log 6 - \frac{2}{3} \log 2) \ .$$

In contrast, from the S/N weighting scheme (3.17):

$$S_{t_1} = \log TOLTF_1 - H(P_{t_1}) = \log 4 - \log 4 = 0 \ ,$$

$$S_{t_2} = \log TOLTF_2 - H(P_{t_2}) = \log 6 - (\log 6 - \frac{2}{3} \log 2) = \frac{2}{3} \log 2 \ ,$$

$$S_{t_3} = \log TOLTF_3 - H(P_{t_3}) = \log 12 - (\log 6 - \frac{2}{3} \log 2) = \frac{5}{3} \log 2 \ .$$

It can be seen from this example that the S/N method has a number of problems. First, whenever a term occurs uniformly in a subset of documents, its value is always equal to zero independent of its document frequency. A second problem is that a term with a large total occurrence frequency will tend to have a higher term weight. These problems are absent in both the IDF and our weighting schemes.

4. Conclusion

By using the concepts developed for evidential reasoning and information theory, we introduce a measure for term weighting. In particular, we demonstrate that the IDF weighting scheme is a special case of the proposed measure. More importantly, in contrast to some earlier explanations for the IDF method, our argument is *not* based on the probabilistic independence model.

Acknowledgment

This paper was prepared while the authors visiting the Department of Computer Science, Cornell University. They wish to thank J. Hopcroft for providing the opportunity to work at Cornell and they are particularly indebted to G. Salton for his encouragement and hospitality.

Reference

[1] Brookes, B.C. "The Shannon model of IR systems," *Journal of Documentation,* **28,** 160-162, 1972.

[2] Chen, Z. "An information theoretic approach to information retrieval," *1988 IEEE International Symposium on Information Theory,* Kobe, Japan, 107-108, 1088.

[3] Croft, W.B. "Document representation in probabilistic models of information retrieval," *Journal of the American Society for Information Science,* **32,** 451-457, 1981.

[4] Croft, W.B., and Harper, D.J. "Using probabilistic models of document retrieval without relevance information," *Journal of Documentation,* **35,** 285-279, 1979.

[5] Dennis, S.F. "The construction of a thesaurus automatically from a sample of text," in Stevens, M.E., Giuliano, V.E., and Heilprin, L.B. (ed.) *Statistical Association Methods for Mechanized Documentation, Symposium Proceedings,* Washington, National Bureau of Standards Miscellaneous Publication 269, 61-148, 1964.

[6] Guiasu, S. *Information Theory with Applications,* New York, McGraw-Hill, 1977.

[7] Hwang, C.L., and Yoon, K. I. "Multiple Attribute Decision Making, Methods and Applications," Berlin, Springer-Verlag, 1981.

[8] Maron, M.E., and Kuhns, J.L. "On relevance, probabilistic indexing and information retrieval," *Journal of the Association for Computing Machinery,* **7,** 216-244, 1960.

[9] Robertson, S.E. "Term specificity," *Journal of Documentation,* **28,** 164-165, 1972.

[10] Robertson, S.E. "Term frequency and term value," *Proceedings of the Fourth ACM SIGIR Conference on Research and Development in Information Retrieval, ACM SIGIR Forum,* **XVI,** 22-29, 1981.

[11] Robertson, S.E., and Sparck Jones, K. "Relevance weighting of search terms," *Journal of the American Society for Information Science,* **27,** 129-146, 1976.

[12] Salton, G. *A Theory of Indexing,* Regional Conference Series in Applied Mathematics, No. 18, Philadelphia, Society for Industrial and Applied Mathematics, 1975.

[13] Salton, G., and Buckley, C. "Term weighting approaches in automatic text retrieval," *Information Processing and Management,* **24,** 513-523, 1988.

[14] Salton, G., and McGill, M.H. *Introduction to Modern Information Retrieval,* New York, McGraw-Hill, 1983.

[15] Shannon, C.E. "A mathematical theory of communication," *Bell System and Technical Journal,* **27,** 379-423, 623-656, 1948.

[16] Sparck Jones, K. "A statistical interpretation of term specificity and its application in retrieval," *Journal of Documentation,* **28,** 11-21, 1972.

[17] Sparck Jones, K. "Experiments in relevance weighting of search terms," *Information Processing and Management,* **15,** 133-144, 1979.

[18] Van Rijsbergen, C.J. *Information Retrieval,* London, Butterworth, 1979.

[19] Wong, S.K.M., and Yao, Y.Y. "Query formulation in linear retrieval models," To appear in *Journal of the American Society for Information Science,* 1988.

[20] Wong, S.K.M., Yao, Y.Y., and Bollmann, P. "Linear structure in information retrieval," *Proceedings of the Eleventh ACM SIGIR Conference on Research and Development in Information Retrieval,* 219-232, 1988.

[21] Wu, H., and Salton, G. "A comparison of search term weighting: term relevance vs. inverse document frequency," *Proceedings of the Fourth ACM SIGIR Conference on Research and Development in Information Retrieval, ACM SIGIR Forum,* **XVI,** 30-39, 1981.

[22] Yu, C.T., and Salton, G. "Precision Weighting -- An Effective Automatic Indexing Method." *Journal of the Association for Computing Machinery,* **23,** 76-88, 1976.

[23] Zeleny, M. *Linear Multiobjective Programming,* New York, Springer-Verlag, 1974.