A Theory of Indexing

G. Salton

Abstract. The content analysis, or indexing problem, is fundamental in information storage and retrieval. Several automatic procedures are examined for the assignment of significance values to the terms, or keywords, identifying the documents of a collection. Good and bad index terms are characterized by objective measures, leading to the conclusion that the best index terms are those with medium document frequency and skewed frequency distributions.

A discrimination value model is introduced which makes it possible to construct effective indexing vocabularies by using phrase and thesaurus transformations to modify poor discriminators—those whose document frequency is too high, or too low—into better discriminators, and hence more useful index terms.

Test results are included which illustrate the effectiveness of the theory.

1. Introduction. Among the various components of a standard information processing environment, the analysis and content identification of the stored records is probably the most crucial one. Indeed, the outcome of the content analysis directly affects the storage organization, search strategy and retrieval properties of the stored information.

Normally, this analysis, or *indexing* operation, consists in the assignment to the stored records of attributes, chosen so as to represent collectively the information content of the corresponding records. Specifically, consider a collection D of stored items D_i. The indexing task then takes on two aspects:

(a) First, it is necessary to choose a set of t distinct attributes A_k which can represent the information content in D.

(b) For each attribute A_k, a number of different values $a_{k_1}, a_{k_2}, \cdots, a_{k_{n_k}}$ are defined, and one of these n_k values is assigned to each record D_i for which attribute A_k applies.

In a file of personnel records, the attributes A_k might be employee name, job classification, department number, salary, and so on. The corresponding attribute-values may be particular names of individual employees, particular job classifications and department numbers, and specific salary levels. The indexing operation then generates for each stored item an *index vector*

$$(1) \qquad D_i = (a_{i1}, a_{i2}, \cdots, a_{it}),$$

where a_{ij} denotes the value of attribute A_j in item D_i. When a given a_{ij} is null, the corresponding attribute is assumed to be absent from the item description. The *attribute-values* a_{ij} are also known as keywords, terms, content identifiers, or simply keys.

A given attribute-value assigned to an item may be weighted by assigning an importance parameter w_{ij} to each a_{ij}, or alternatively it may be unweighted. In the latter case, the weights w_{ij} are restricted to the values 0 or 1, a 1 being automatically assigned as the weight of each keyword present in, or applicable to, a given index vector, and a 0 to each keyword that is not applicable. Unweighted index vectors are also known as binary, or logical vectors.

In principle, a complete index vector then consists of sets of pairs (a_{ij}, w_{ij}) as follows:

$$D_i = (a_{i1}, w_{i1}; a_{i2}, w_{i2}; \cdots; a_{it}, w_{it}), \tag{2}$$

where w_{ij} denotes the weight of term a_{ij}. In practice, one can avoid storing either the keywords or the weights in one of two different ways. When the vectors are binary, the vector elements may be restricted to include only those keywords whose weight equals 1 by eliminating terms of 0 weight; obviously, the weight indications are then redundant.

Alternatively, when the number of possible attribute-values is limited, a fixed position may be assigned to each attribute-value in the index vector. In that case, the weights alone suffice to specify the index vectors, a zero weight being used to identify keys that do not apply to a given item.[1] In that system, the vector (0, 0, 0, 15, 0, 0, 5, 0) might then denote the presence of terms 4 and 7 with weights 15 and 5, respectively.

Given an indexed collection, it is possible to compute a similarity measure between pairs of items by comparing the corresponding vector pairs. A typical measure of similarity s between items D_i and D_j might be

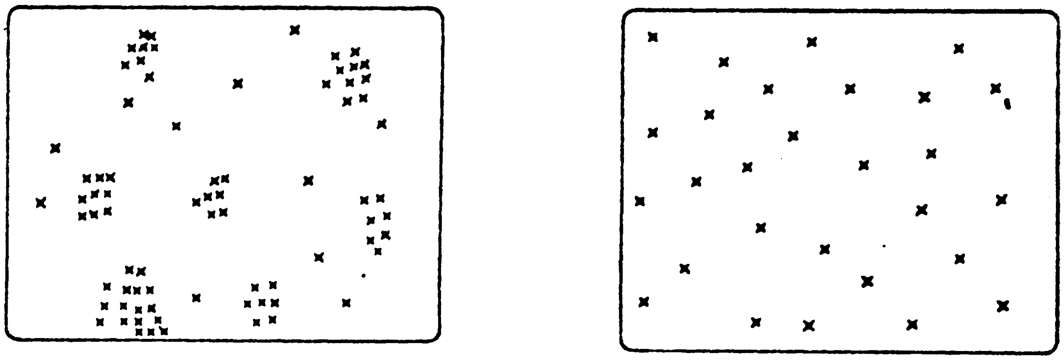
$$s(D_i, D_j) = \sum_{k=1}^{t} w_{ik} w_{jk}. \tag{3}$$

For binary vectors, this equals the number of matching keywords in the two vectors, whereas for weighted vectors it is the sum of the products of corresponding term weights.

In some indexing systems, additional relations are defined between certain attributes or attribute-values included in the index vectors. In that case, appropriate relational indicators must be included in the index vectors; the vector images may then be transformed into graphs, each node of the graph representing a keyword, and the labelled branches between pairs of nodes specifying the relations. The computation of the similarity between two items is then transformed into a graph matching process, where nodes (keywords) are compared as well as branches (relations between keywords).

No matter what particular indexing system is used, an effective indexing vocabulary will produce a clustered object space in which classes of similar items are easily separable from the remaining items. A typical example is shown in Fig. 1(a), where a cross (×) denotes each item, and the distance between two items is inversely proportional to the similarity of their index vectors. Obviously, when the

[1] In practice, most keys will be absent from most index vectors; instead of storing the resulting sparse vectors directly, a compression scheme may be used to delete the large number of zeros, while still allowing proper decoding of the stored information.



(a) Clustered object space (b) Object space lacking separation

FIG. 1. *Typical object space configurations*

object space configuration is similar to that shown in Fig. 1(a), the retrieval of a given item will lead to the retrieval of many similar items in its vicinity, thus ensuring high recall; at the same time extraneous items located at a greater distance are easy to reject, leading to high precision.[2] On the other hand, when the indexing in use leads to an even distribution of objects across the index space, as shown in Fig. 1(b), the separation of relevant from nonrelevant items is much harder to effect, and the retrieval results are likely to be inferior.

It would be nice to relate the properties of a given indexing vocabulary directly to the clustering properties of the corresponding object space. Unfortunately, not enough is known so far about the relationship between indexing and classification to be precise on that score. The properties of normal indexing vocabularies are related instead to concepts such as *specificity* and *exhaustivity*, where term specificity denotes the level of detail at which concepts are represented in the vectors, whereas the indexing exhaustivity designates the completeness with which the relevant topic classes are represented in the indexing vocabulary. The implication is that specific index vocabularies lead to high precision searches (that is, to the rejection of nonrelevant materials), whereas exhaustive object descriptions lead to high recall.

In principle, exhaustivity and specificity are independent properties of the indexing environment. In practice, exhaustive indexing products are easier to generate using broad (nonspecific) index terms, and contrariwise, the use of highly specific terms often leads to insufficiently exhaustive index vectors. This phenomenon explains in part the well-known inverse relation between recall and precision: searches can be conducted so as to produce high recall (the retrieval of much relevant material), generally at the cost of low precision (the retrieval of much extraneous material at the same time); contrariwise high precision normally entails low recall.

Attempts have been made to relate standard parameters such as exhaustivity and specificity to quantitative measures, including the length of the indexing

[2] Recall is the proportion of relevant items retrieved, while precision is the proportion of retrieved items that are relevant. Normally, most relevant items should be retrieved, while most nonrelevant should be rejected, leading to high recall, as well as high precision.

vectors (number of terms included in a vector) representing exhaustivity, and the number of distinct vectors to which a term is assigned to denote inverse specificity [1], [2]. Such formal characterizations may in time lead to the use of optimal indexing vocabularies and the construction of optimal indexing spaces. These questions are considered in the remainder of this study.

2. Term significance computations.

A. *Term frequency parameters.* Most automatic indexing experiments have been conducted in library or information center environments. In that case, the vectors represent documents, or other information items, and the terms are subject identifiers representing document content. There is agreement that the original document, or at least some document excerpt such as a title or abstract, must form the basis for the initial indexing. Furthermore, special provisions are always made for high-frequency common function words, such as "of", "and", "but", etc. Normally, they are simply deleted by referring to a so-called "stop" list, containing terms chosen for elimination.

Beyond this, a great variety of different practices have come to be implemented, all of them designed to lead to the construction of good indexing vocabularies. The simplest possible indexing process consists in the assignment of an importance factor (weight) to each word extracted from a document excerpt, followed by the inclusion of highly weighted terms in the indexing vectors of the corresponding document vectors. This method stands, or falls, with the choice of a good weighting function.

The best known of these functions are the basic frequency measures originally introduced by Luhn [3], [4], including in particular the *term frequency*, that is, frequency of occurrence of term k in the ith document f_i^k, as well as the total *collection frequency* F^k of term k, defined for n documents as

$$F^k = \sum_{i=1}^{n} f_i^k. \tag{4}$$

When the term frequency f_i^k or the collection frequency F^k is used as an indicator of term importance, those terms which occur most often in the collection, or in the individual documents, are assumed to be the most valuable terms. While high-frequency terms are likely to produce a large number of matches between query and document vector elements and lead to the retrieval of many relevant documents, the usefulness of term and collection frequency weights may be questioned on information theoretic grounds [5]. In particular, the frequent terms—those assigned to a large proportion of the documents in a collection—carry relatively less information than the rarer terms, and they may not be effective in distinguishing the relevant from the nonrelevant items.

These considerations lead to the notion that the best terms should be those which are emphasized in certain specific items in the collection, while over the whole collection their occurrence frequencies are generally low. A possible measure of the importance of term k in document i would then be f_i^k/F^k. Alternatively, another frequency-based parameter may be introduced as the *document frequency* B^k, where

$$B^k = \sum_{i=1}^{n} b_i^k, \tag{5}$$

and $b_i^k = 1$ whenever $f_i^k > 0$, and $b_i^k = 0$ otherwise. B_i^k then represents the number of documents in which term k occurs, an appropriate term weighting function being f_i^k/B^k.

Still another possibility consists in emphasizing those terms which are highly weighted in particular document collections, while being of relatively small importance in the literature-at-large [6]. Such relative frequency parameters are, however, difficult to utilize because the "literature-at-large" cannot easily be captured.

B. *Signal-noise parameters*. The frequency parameters introduced in the previous subsection measure the importance of a given term by its frequency in individual documents, possibly supplemented by total collection or document frequency counts. A more complete picture of term behavior may be obtained by considering the frequency characteristics of each term not only in the particular document whose term weights are currently under assessment, but also in all other documents in a collection. One such measure also derived from communications theory is the so-called *signal-noise* ratio [5], [7]. **Specifically, for a collection of n** documents, the noise N^k of term k is defined as

$$N^k = \sum_{i=1}^{n} \frac{f_i^k}{F^k} \log \frac{F^k}{f_i^k}, \tag{6}$$

and the signal S^k is correspondingly

$$S^k = \log F^k - N^k. \tag{7}$$

The noise N^k is a function of the evenness of the frequency distribution of term k among the documents in which term k appears. Alternatively, the noise may be said to vary inversely with the "concentration" of a term in the document collection. For perfectly even distributions, a term occurs an identical number of times in each document of the collection. In these circumstances, the noise will be maximized.

Consider, for example, the case where term k occurs exactly once in each document (all $f_i^k = 1$). In that case, $F^k = n$ and

$$N^k = \sum_{i=1}^{n} \frac{1}{n} \log \frac{n}{1} = \log n.$$

Obviously, a zero signal is produced in that case. On the other hand, for perfectly concentrated distributions, a term will appear in only one document of a collection with frequency F^k. The noise will then be zero, and the signal optimum, because

$$N^k = \frac{F^k}{F^k} \log \frac{F^k}{F^k} = \log 1 = 0$$

and

$$S^k = \log F^k - N^k = \log F^k.$$

The relation of equation (7) makes it clear that high noise implies low signal, and vice versa. A relation also exists between noise and term specificity, and between signal and total collection frequency of a term. In general, broad, nonspecific terms tend to have more even distributions, hence high noise, while high document frequencies may also produce large signals. These relations are, however, only approximate for high-frequency terms which also exhibit even distributions, since the noise is then also substantial. Possible weighting functions based on the signal-noise parameter may be S^k/N^k, or alternatively $(S^k/N^k) \cdot S^k$ (see [7]).

Signal-noise computations may be used to construct an optimal indexing vocabulary by deleting terms which exhibit excessively low signal-noise values [7]. In particular, consider a figure of merit for the m terms used to index a given document collection, such as

$$FM = \frac{\sum_{i=1}^{m} S^i}{\sum_{i=1}^{m} N^i} \cdot \sum_{i=1}^{m} S^i. \tag{8}$$

If FM_j is the figure of merit with term j deleted, that is,

$$FM_l = \frac{\sum_{\substack{i=1\\ i\neq l}}^{m} S^i}{\sum_{\substack{i=1\\ i\neq l}}^{m} N^i} \cdot \sum_{\substack{i=1\\ i\neq l}} S^i,$$

then an optimal vocabulary may be obtained by deleting terms l so as to maximize the function $FM_l - FM$. Indeed, consider a term j for which N^j is very large, while S^j is small. The removal of such a term will ensure that $FM^j > FM$, and the difference in the figures of merit will grow.

When the terms in a collection are ordered by a parameter proportional to the signal to noise ratio, it develops that the best signal-to-noise terms have low overall document frequency and concentrated distributions; bad signal-to-noise terms also have low document frequency but even distributions (they occur in many documents).

The signal-to-noise ratio S^k/N^k can be used directly to obtain a global weighting factor for each term in a collection, leading to the deletion of terms with insufficient S/N ratios. To obtain term weights valid for a given term in a specific document, the S/N ratio may be combined with the term-frequency parameters previously described. A possible value for term k in document i might then be $(f_i^k/F^k)(S^k/N^k)$. Such functions are examined again later in this study.

C. *Parameters based on variance.* The variance V^k of the term frequencies for term k is defined as

$$(V^k)^2 = \frac{\sum_{i=1}^{n} (f_i^k - \bar{f}^k)^2}{n-1}, \tag{9}$$

where n is the number of documents in the collection, and $\bar{f}^k$ is the average term frequency for term k across the n documents, that is, $\bar{f}^k = F^k/n$. Obviously, the variance will be small for terms exhibiting even frequency distributions (all f_i^k are approximately equal to $\bar{f}^k$), and for terms which occur in very few documents (most f_i^k are equal to zero, and $\bar{f}^k$ is near zero). On the other hand, when a term exhibits a skewed distribution, and at least medium collection frequency F^k, then the variance may be large.

The use of term importance parameters which are based on the variance of the frequency distribution may be justified by the notion that good terms must necessarily be able to distinguish the various documents from each other. This eliminates terms with even frequency distributions and low variance, and favors those with large variations in the individual term frequencies, and hence high variance.

Among the various measures that are based on the variance of the term frequency distribution, the most satisfactory is the one called *NOCC/EK* by Dennis, or *EK* for short [8]. It varies directly with the variance, and inversely with the collection frequency F^k, thus again giving preference to the rarer terms among those with high variance. The following formula can be used for the computations:

$$NOCC/EK = (EK)_k = \frac{F^k(V^k)^2}{(\bar{f}^k)^2}. \tag{10}$$

Replacing $\bar{f}$ by F/n, and using a denominator equal to n, instead of $n-1$, in the variance formula (9), one obtains

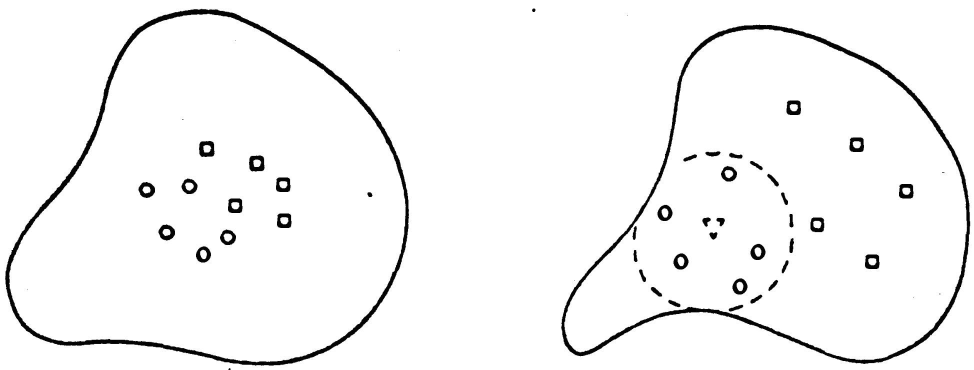
$$(EK)_k = \frac{F^k(V^k)^2}{(\bar{f}^k)^2} = \frac{F^k}{(\bar{f}^k)^2 n} \sum_{i=1}^{n} (f_i^k - \bar{f}^k)^2$$

$$= \frac{n}{F^k} \sum_{i=1}^{n} (f_i^k)^2 - F^k. \tag{11}$$

The expression of formula (11) shows that the variance measure is even more sensitive to large individual term frequencies than the previous measures. The best *EK* terms are those whose collection frequency F^k is not too large, and whose frequency distribution is concentrated so as to produce a large sum for the f_i^k terms. The worst *EK* terms are those with a large collection frequency F^k and even term distributions.

As for the signal-noise ratio, the *EK* parameter assigns a global value to each term in a collection. For document indexing purposes, it must be supplemented by local term values valid within each document alone. A possible weight for term k in document i might then be $(f_i^k/F^k) \cdot (EK)_k$.

D. *Parameters based on discrimination values.* The discrimination value model rates the potential index terms in accordance with their usefulness as document discriminators: in addition, it offers the advantage of providing a reasonable physical interpretation for the indexing process [9], [10]. Specifically, the assumption is that a document space which is "bunched up" in the sense that all documents exhibit somewhat similar index vectors is not useful for retrieval, since one document cannot then be distinguished from another: contrariwise, a space which is spread out in such a way that the documents are widely separated from each other provides an ideal retrieval situation, since some documents may then be retrieved, while others can be rejected. A typical document environment is represented in Fig. 2, where, once again, the distance between two items is inversely related to the similarity of their index vectors. In the example of Fig. 2(a), little separation is provided between the set of relevant and nonrelevant items; in Fig. 2(b), on the other hand, which is produced by the incorporation of discriminating terms into the document vectors, the query construction and retrieval tasks appear much easier to perform.



(a) *Original space*

(b) *Space "spread out" by assigning discriminators*

FIG. 2. *Term discrimination model.* □ *nonrelevant document*; ○ *relevant document*; ▽ *query*; ⁐ *retrieval region*.

The discrimination value model leads to a distinction among possible index terms in accordance with their ability to "spread out" the document space when assigned to the documents of a collection.

Consider a collection of n documents $\{D\}$, and let each document D_i be identified by vector elements $w_{i1}, w_{i2}, \cdots, w_{it}$ as before. Let $s(D_i, D_j)$ represent the similarity between documents i and j, measured by a comparison between the corresponding document vectors. If the measure s is computed for all pairs of items (D_i, D_j) such that $i \neq j$, an average value $\bar{s}$ can be produced representing the average document pair similarity for the collection. Specifically,

$$\bar{s} = K \sum_{\substack{i=1, j=1 \\ i \neq j}}^{n} s(D_i, D_j) \tag{12}$$

with K constant. Obviously, the value of $\bar{s}$ represents a measure of space density, since a large $\bar{s}$ identifies a "bunched up" environment with large average document pair similarities, whereas a small $\bar{s}$ implies that the space is spread out.

Consider now the original document collection with term k removed from all the document descriptions and let $\bar{s}_k$ represent the average document pair similarity in that case. If term k represents a broad, high-frequency term with a fairly even frequency distribution, it is likely that it would have appeared in most document descriptions; its removal from the individual document vectors will therefore decrease the average document pair similarity, so that $\bar{s}_k < \bar{s}$. Contrariwise, when term k exhibits a skewed distribution, in the sense that it occurs with high weight in some document vectors but not in others, its removal is likely to increase the average term pair similarity (since its assignment reduces that same similarity), or $\bar{s}_k > \bar{s}$.

A *discrimination value* can now be computed for each term k, as a function of the value $(\bar{s}_k - \bar{s})$ which assigns positive weights to the good discriminators—those causing an increase in document-pair similarity when removed (or a decrease when assigned)--and negative ones to the bad discriminators. The terms can then be arranged in decreasing order in accordance with the discrimination value, and a discrimination value weighting system can be used to emphasize good discriminators and deemphasize the poor ones. If $(DV)_k$ is the discrimination value of term k, a possible weighting function for term k in document i might be $(f_i^k\ F^k)\cdot(DV)_k$.

In practice the computation of average document pair similarities $\bar{s}$ and $\bar{s}_k$ requires of the order of $(t + 1)n(n - 1)/2$ vector comparisons for n documents and t terms. This can be reduced to $(t + 1)n$ comparisons by introducing a central item, or centroid C, of the document space, representing the average document, where the ith vector element c_i is defined as

$$c_i = \frac{1}{n}\sum_{j=1}^{n} w_{ij},$$

that is, as the average weight of term i in all n documents. This leads to a space density function Q defined simply as the sum of the similarity coefficients between centroid C and all documents D_i, that is,

$$Q = \sum_{j=1}^{n} s(C, D_j). \tag{13}$$

When $0 \leqq s \leqq 1$, then $0 \leqq Q \leqq n$.

If Q_k represents the space density Q of expression (13) with term k removed from all document vectors, the discrimination value $(DV)_k$ for term k may then be defined simply as $Q_k - Q$. Obviously, for good discriminators $Q_k - Q$ is positive, because the removal of term k will cause the space to become more dense; hence $Q_k > Q$. For poor discriminators the reverse obtains.

Figure 3 illustrates the situation where a discriminator is removed from the document vectors; the similarity between most items and the centroid becomes larger (the distances are reduced between corresponding points), and the space density increases.

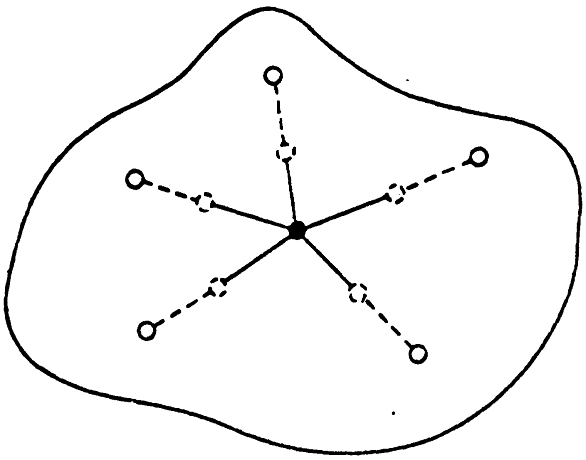


FIG. 3. *Discrimination value computation ($Q_k > Q$). ● space centroid; ○ original documents; ◌ documents following removal of discriminator.*

When the terms are arranged in decreasing order according to the $Q_k - Q$ function, it is found that best terms have average document frequency—neither too high nor too low—and frequency distributions that are fairly skewed. Bad discriminators, on the other hand have high collection frequency, and are present in most documents of a collection. Average discrimination values are obtained for very low frequency terms. These characterizations are useful to derive an appropriate indexing theory, as shown later in this study.

E. *Parameters based on dynamic information values.* The term significance calculations based on the use of *dynamic information values* are different from all others, in that the term values are not primarily derived from collection-dependent properties. Instead, the terms occurring in a collection of documents may all be equally weighted initially, for example by being assigned a common average weight. A weight adjusting process can then be used to promote some terms by increasing their weight, while similarly demoting others. The terms chosen for promotion are often those for which some positive information is available—for example, they may be assigned to retrieved documents identified as relevant by the user population in the course of a retrieval operation. The demoted terms may similarly be those occurring in nonrelevant documents that may be retrieved.

A particular form of dynamic information value, due to Sage, Anderson and Fitzwater, specifies starting values equal to 1, which can successively be adjusted upward to 2, or downward to 0, depending on the term occurrence properties—that is, on their inclusion in retrieved items that may be either relevant or nonrelevant [11]. The alteration process is performed in such a way that terms in the middle of the weight range, where the values are close to 1, are shifted more rapidly than those near the edges of the range (that is, close to 0 or 2), the hope being that equilibrium values for the terms can then be achieved more rapidly. Specifically, a transformation is used through a sine function, which produces larger differences in functional values near $x = 0$, than near $x = \pi/2$, or $x = -\pi/2$.

Consider the following definitions: Let

$$v_i = \text{information value of term } i \qquad 0 \leq v_i \leq 2,$$
(initially all $v_i = 1$),

$$x_i = \arcsin(v_i - 1)$$
the transposed information value.

Then

$$v_i = 1 + \sin x_i.$$

A value of Δx is then chosen as a function of the existing information value, where

$$\Delta x = \frac{\pi/2 - |x_i|}{c}, \qquad c = \text{const}.$$

This gives rise to a new, updated information value

(14) $$v_i^* = 1 + \sin(x_i \mp \Delta x_i).$$

In the updating process, the + sign obtains when the term must be promoted, or increased in value—for example, when in a retrieval environment a query term happens to be present in a retrieved document identified as relevant by the user population; in the opposite case, the minus sign obtains. A graphic representation of the term adjustment process is included in Fig. 4.

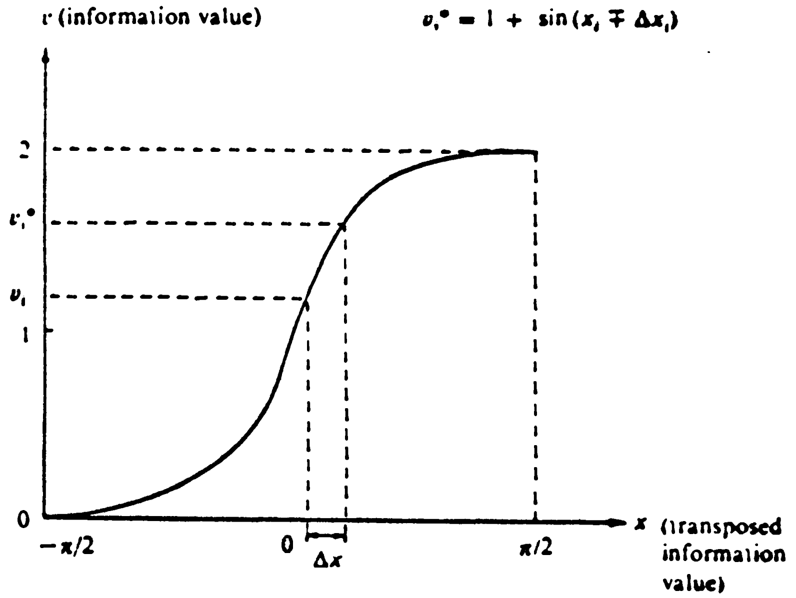


FIG. 4. *Information value construction.*

It has been stated that the dynamic term adjustment process will converge to some optimum value for each term, since false high weights will lead to the retrieval of nonrelevant items, thus eventually producing weight reductions, whereas false low weights will similarly produce an upward adjustment of term weights.

The five parameter types described in this section all respond to different criteria of importance, and there may in fact be no one algorithm that would be optimal for all indexing situations. Thus, very low frequency terms which are often thought to be only marginally useful in retrieval (since they produce so few matches between the query statements and the documents) might in fact be given a very high weight—as in the signal-noise ratio—if high precision output were of overriding importance. Similarly, very high frequency terms with low discrimination values might in fact be important when the user insists on high-recall.

The usefulness of one or another of the term significance measures must then depend on the environment under consideration and on the particular user requirements. The same is true of some of the additional text-based criteria that have been used in the past in evaluating individual term importance, such as, for example, word position in the paragraph structure of a given text (words appearing in titles or section headings may be weighted more highly than those appearing in the body of a text), the presence or absence of special indicator words in the immediate context of the given term, the word distance between terms, and so on.

An evaluation of the main term significance measures is included later in this study.

3. Utilization of term significance. The term significance measures previously described are useful for a variety of different purposes. First, and most importantly, the weighted vectors make possible a detailed identification of the objects under consideration. This implies that the similarity between two items can be determined more precisely than would be the case when binary index vectors are used with weights restricted to 0 and 1. Thus, a similarity computation such as that of equation (3) produces simply the number of matching terms when the vectors D_i and D_j are binary; a more complicated function results for weighted vectors.

In a retrieval situation, it becomes necessary to assess the similarity between documents and queries before retrieving items with sufficiently large similarity coefficients. When weighted document and query vectors are used, it is then likely that $s(Q, D_i) \neq s(Q, D_j)$, for all queries Q and documents D_i and D_j, such that $i \neq j$. An ordering of the output documents in decreasing query-document similarity order then produces a strict ranking of the items which can be used to limit the size of the retrieved set to those items which are most likely to be of interest to the user population. A typical ranked output list is shown in Table 1 (from [12, Chap. 1]).

It has been shown that the use of ranked document output considerably enhances the retrieval effectiveness, particularly in those situations where a series of partial searches is used to approach a given topic area little by little. In such cases, feedback information derived from previous search output is often used to construct new, improved query formulations. When these new formulations are based on the top few documents retrieved in a previous search—that is, on those whose

TABLE 1

Retrieval output in decreasing query-document similarity order (adapted from [12])

Rank	Document number	Query-document similarity coefficient
1	384	0.6676
2	360	0.5758
3	200	0.5664
4	392	0.5508
5	386	0.5484
6	103	0.5445
7	85	0.4511
8	192	0.4106
9	102	0.3987
10	358	0.3986
11	387	0.3968
12	202	0.3907
13	229	0.3506
14	88	0.3452
15	251	0.3329

similarity coefficients with the queries are highest—it is often possible to obtain excellent retrieval results in very few search interations [13].

In addition to providing ranked retrieval output, the term significance values can be used to generate associations between terms leading to improved recall by means of the so-called *associative indexing* technique [14]–[16]. The idea is to use similarities between index terms as a basis for defining for each original index term a set of associated terms that can be added to the index vectors, thereby supplying additional search terms.

Most associative indexing methods are based on a prior availability of a *term association matrix* specifying for each term pair the corresponding strength of association. Association factors which exceed in magnitude a predetermined threshold are then assumed to identify term pairs that exhibit a sufficiently high degree of association to be useful for associative indexing purposes. For a collection of n documents, a typical association factor between terms j and k might simply be the sum over all documents of the product of the corresponding term frequencies:

$$f(j,k) = \sum_{i=1}^{n} f_i^j \cdot f_i^k. \tag{15}$$

Alternatively, the association factors might be normalized to produce a coefficient ranging from 0 for perfectly disassociated pairs to 1 for perfectly associated ones. A typical normalized association coefficient is

$$f(j,k) = \frac{\sum_{i=1}^{n} f_i^j \cdot f_i^k}{\sum_{i=1}^{n} (f_i^j)^2 + \sum_{i=1}^{n} (f_i^k)^2 - \sum_{i=1}^{n} f_i^j \cdot f_i^k}. \tag{16}$$

Consider, as an example, the typical term association matrix $\mathbf{D}$ represented in Fig. 5 for the five terms A, B, C, D, and E. If $\mathbf{q}$ is a typical term vector (for example, a query vector), then a new expanded vector $\mathbf{q}'$ may be obtained simply by the

$$\begin{array}{c c} & \begin{array}{ccccc} A & B & C & D & E \end{array} \\ \begin{array}{c} A \\ B \\ C \\ D \\ E \end{array} & \begin{pmatrix} 1 & \frac{1}{2} & \frac{1}{4} & 0 & \frac{1}{4} \\ \frac{1}{2} & 1 & \frac{1}{4} & \frac{1}{2} & 0 \\ \frac{1}{4} & \frac{1}{4} & 1 & \frac{1}{4} & \frac{1}{2} \\ 0 & \frac{1}{2} & \frac{1}{4} & 1 & \frac{1}{2} \\ \frac{1}{4} & 0 & \frac{1}{2} & \frac{1}{2} & 1 \end{pmatrix} = \mathbf{D} \end{array}$$

FIG. 5. *Typical term association matrix* $\mathbf{D}$.

vector equation $\mathbf{D} \cdot \mathbf{q} = \mathbf{q}'$, as shown in Fig. 6. This transforms the original vector $\mathbf{q} = (4, 2, 1, 1, 0)$ into $\mathbf{q}' = (5\frac{1}{4}, 4\frac{3}{4}, 2\frac{3}{4}, 2\frac{1}{4}, 2)$. Thus term A with an original weight of 4 is raised to $5\frac{1}{4}$ by addition of 1 $(2 \cdot \frac{1}{2})$ from the associated term B, plus $\frac{1}{4}$ $(1 \cdot \frac{1}{4})$ from term C. The other weights are altered in a similar manner, as shown in detail in Fig. 6.

$$\underbrace{\begin{pmatrix} 1 & \frac{1}{2} & \frac{1}{4} & 0 & \frac{1}{4} \\ \frac{1}{2} & 1 & \frac{1}{4} & \frac{1}{2} & 0 \\ \frac{1}{4} & \frac{1}{4} & 1 & \frac{1}{4} & \frac{1}{2} \\ 0 & \frac{1}{2} & \frac{1}{4} & 1 & \frac{1}{2} \\ \frac{1}{4} & 0 & \frac{1}{2} & \frac{1}{2} & 1 \end{pmatrix}}_{\mathbf{D}} \underbrace{\begin{pmatrix} 4 \\ 2 \\ 1 \\ 1 \\ 0 \end{pmatrix}}_{\mathbf{q}} = \underbrace{\begin{pmatrix} 4 + 1 + \frac{1}{4} \\ 2 + 2 + \frac{1}{4} + \frac{1}{2} \\ 1 + \frac{1}{2} + 1 + \frac{1}{4} \\ 1 + \frac{1}{4} + 1 \\ 1 + \frac{1}{2} + \frac{1}{2} \end{pmatrix} = \begin{pmatrix} 5\frac{1}{4} \\ 4\frac{3}{4} \\ 2\frac{3}{4} \\ 2\frac{1}{4} \\ 2 \end{pmatrix}}_{\mathbf{q}'}$$

FIG. 6. *Typical associative indexing strategy* ($\mathbf{q}' = \mathbf{D} \cdot \mathbf{q}$).

Many alternative strategies are possible, including for example the use of higher order term associations (see [12, Chap. 4]). Thus if term A is associated with B, and B is associated with C, a second order association exists between A and C; if in addition C is also associated with D, then a third order association may be defined between A and D. In practice, higher order associations are not likely to be used, first, because of the increasingly more expensive computations needed to perform the necessary processing—even first order associations require t^2 operations to generate the association matrix for t terms, and second, because of the small likelihood of determining useful relations in this manner.

A process somewhat similar to associative indexing is the so-called *probabilistic indexing*, in which the presence of certain terms in the documents is used as a criterion for the assignment to the documents of additional class identifiers [17], [18]. These class identifiers then play the role of the recall-enhancing associated terms previously discussed. Specifically, the assignment of terms $T_1, T_2, \cdots, T_i$ to document D_j is used as a basis for stating that document D_j belongs to category C_k with probability p. When p is large enough, D_j is assigned to C_k, and the corresponding class identifier can be added to the set of terms identifying the document.

The actual computations are performed by noting that when the terms are independently assigned, the probability of class k obtaining, given terms $T_1, T_2, \cdots, T_i$, equals the a priori probability of class C_k, multiplied by the individual probabilities that an item in class C_k will individually contain each of the terms $T_1, T_2, \cdots$, up to T_i. That is,

$$P(T_1, T_2, \cdots, T_i, C_k) = a \cdot P(C_k) \cdot P(C_k, T_1) \cdot P(C_k, T_2) \cdots \cdot P(C_k, T_i).$$

The constant a is so chosen that the total probability of assignment of a given document to all m classes equals 1, or

$$\sum_{k=1}^{m} P(T_1, T_2, \cdots, T_i, C_k) = 1,$$

thus implying that the subject classes are mutually exclusive and exhaustive (that is, that each document belongs to one and only one class).

It remains to show how to estimate the a priori class probabilities $P(C_k)$, and the joint probabilities $P(C_k, T_i)$ which specify the likelihood that if item D_j is in class C_k, it will contain term T_i. An easy way of doing this is to use statistical information derived from the class assignments and term weights of an existing document collection as follows:

$P(C_k)$ is approximated by taking the total number of document assignments to class C_k divided by the number of document assignments to all m classes; and

$P(C_k, T_i)$ is assumed to be the total number of occurrences of the sum of the weights of term T_i in documents assigned to class C_k, divided by the total number of term occurrences or the total weights for all t terms for documents in class C_k.

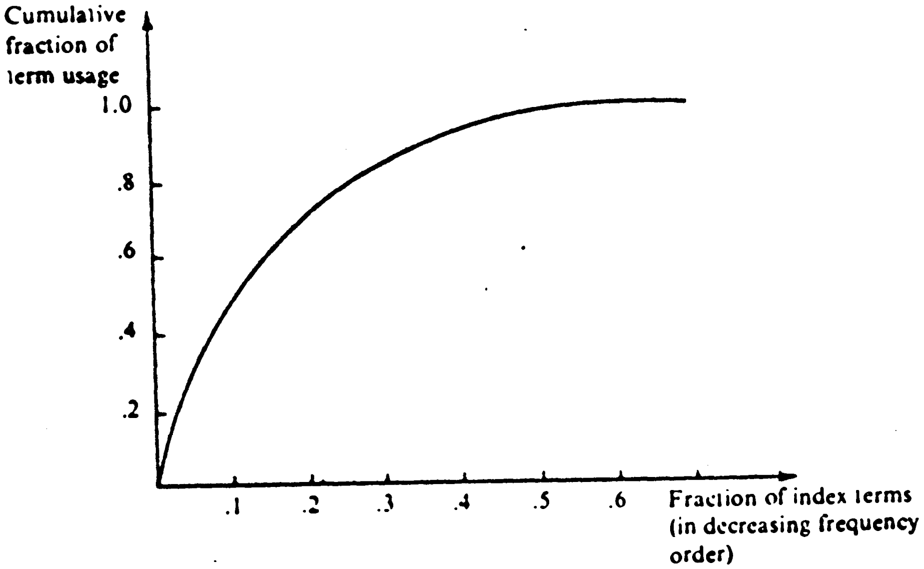
Although the foregoing methodology is based on a number of simplifying assumptions that are untenable in practice – for example, terms are not normally independently assigned to documents, and class assignments are not usually mutually exclusive – it has been shown experimentally that when a sufficient number of terms is available for document identification, the "correct" class C_k can be determined with probabilities ranging from 85 to 100 percent [18].

Possibly the most important application of the term significance computations relates to the specification of an indexing vocabulary of optimum size. There is agreement that an effective indexing vocabulary must include some general terms that can retrieve a large number of relevant documents thereby enhancing the recall; if high precision searches are to be made possible at the same time, some specific terms are needed also in order to make possible an accurate retrieval of individual relevant documents.

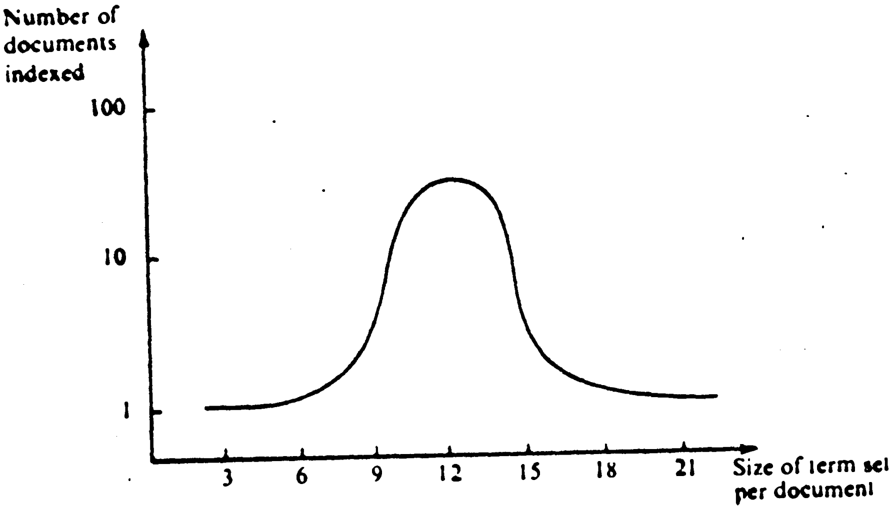
These considerations do not unfortunately lead directly to the determination of good, or bad index terms. This question is normally approached by performing a study of existing indexing vocabularies in order to determine the appropriate occurrence characteristics and frequency distributions. A number of patterns appear to emerge:

(a) In general, a small number of heavily used index terms accounts for a large proportion of index term usage; typically, the most used twenty percent of the terms may constitute sixty to seventy percent of the total term assignments to the documents of a collection. A typical curve showing the fraction of index terms against cumulated term usage is included in Fig. 7(a) (see [19], [20]).

(b) When the length of the indexing vectors is considered, that is, the number of terms assigned to individual documents, the distribution is often log-normal.



(a) *Index term usage statistics*



(b) *Term assignment to documents*

FIG. 7. *Term frequency characteristics.*

Specifically, the number of terms per document appears to be normally distributed about the mean when plotted against the logarithm of the number of documents, as shown in the example of Fig. 7(b) (see [21], [22]).

(c) The growth of the indexing vocabulary as a function of collection size appears to follow empirical laws such as

(17) $$t = a \log (n + b) - c,$$

where t and n are the sizes of the term and document sets, respectively, and a, b and c are constants [21].

While none of these observations can be translated directly into the choice of an appropriate indexing vocabulary, the term significance measures might be used immediately to reduce the size of an existing vocabulary to some optimum value related to collection size—for example, by using equation (17) as a guide—by eliminating terms exhibiting low significance values. More generally, information about the ideal size of a given indexing vocabulary and about the distribution of the vector length of typical index vectors representing document content (points (a), (b) and (c) above) might be combined with the term significance computations to generate ideal indexing vectors exhibiting appropriate length and distribution characteristics and high information content [22], [23]. Attempts at generating an indexing theory including a variety of the previously mentioned models are described later in this study.

4. Characterization of term significance rankings. Before presenting some of the experimental evidence pertaining to the use of term significance computations, it may be of interest to characterize the terms classified as good, average, or poor, respectively, according to the five significance measures previously introduced, including discrimination values (DV), inverse document frequencies ($1/B$), signal-noise values (S/N), variance-based measures (EK), and information values (IV).

A list of terms obtained from a collection of 425 documents in world affairs is shown in Table 2 arranged in ranked order according to four of the significance measures, including DV, S/N, EK, and $1/B$. The 15 best and 15 worst terms are shown in each case chosen from a vocabulary of 7569 terms in world affairs. Entries are not included for the information value rankings because in the laboratory it is difficult to produce a stable set of information values with the limited term value alterations occurring in the experimental situation.

An examination of the terms included in Table 2 shows that the entries occupying the top 15 ranks are all specific topic indicators; the terms at the bottom of the list, on the other hand, are of a more general nature and include elements which are obviously not suitable for content identification. Some overlap is seen to exist between the top discriminators, and the signal-noise, and EK terms. In general, however, the lists are substantially different.

Of the four significance methods illustrated in Table 2, a ranking useful for retrieval purposes is not obtained when the terms are arranged in inverse document frequency order. Indeed, the top of the list is then occupied by several dozen, or even hundreds, of terms with document frequency B^k equal to 1. Obviously such terms are only marginally useful in retrieval because of their excessive rarity. Typical term frequency distributions for three categories of terms in inverse document frequency order are shown in Table 3 for a collection of 200 documents in aerodynamics. It may be seen that the terms with low ranks and hence high values have uninteresting distributions. On the other hand, the terms with ranks 734 to 736 which occur in about half of the items in the collection exhibit less uniform frequency distributions. These terms may in fact be useful in retrieval, although they are assigned low ranks, using the I B procedure.

A detailed examination of the remaining three ranking systems, including *DV*, *S/N*, and *EK* is included in Tables 4 and 5. Consider first the output of Table 4

TABLE 2
Fifteen best and worst terms using four term significance measures 1425 articles in world affairs from Time)

Rank	Discrimination value	Signal Noise	EK Value	Inverse document frequency I B*
1.	Buddhist	Irish	Irish	Amalt
2.	Diem	Ireland	Ireland	Quinim
3.	Lao	Lemass	Lemass	Cynthia
4.	Arab	Dublin	Nasser	Shakhbut
5.	Viet	Rachman	Malay	Fraternity
6.	Kurd	Wynne	Kurd	Roberto
7.	Wilson	Kurd	Arab	Petra
8.	Baath	Liechtenstein	Tunku	Marj
9.	Park	Schweitzer	Chin	Sobukwe
10.	Nenni	Krim	Minh	Dolci
11.	Labor	Zermatt	Dublin	Swan
12.	Macmillan	Ching-Kuo	Rachman	Kaunda
13.	Hassan	Malay	Wynne	Script
14.	Tshombe	Argond	Baath	Brickbats
15.	Nasser	Amah	Buddhist	Vaduz
.	.	.	.	.
7555.	Count	Brief	Insist	Official
7556.	War	Crack	Look	Arm
7557.	West	Purpose	Worse	Work
7558.	Arm	One time	Swept	Stateless
7559.	Force	Bitterly	Prepare	Count
7560.	Work	Kind	Brief	War
7561.	Lead	Huge	Crack	Force
7562.	Red	Insist	Purpose	Minister
7563.	Minister	Taking	One time	Party
7564.	Nation's	Doing	Bitterly	Lead
7565.	Party	Discover	Doing	U.S.
7566.	Commune	Prepare	Discover	Commune
7567.	U.S.	Indeed	Indeed	Nation's
7568.	Govern	Alone	Alone	Govern
7569.	New	Shot	Shot	New

* Top 15 in column 4 chosen randomly from those terms with document frequency of one.

TABLE 3

Frequency distribution of sample terms in inverse document frequency (1/B) order (CRAN 200 collection — 736 term classes)

Characterization	Term number	Rank	F^k	B^k	Number of documents in which term appears with f_i^k of: 1	2	3	4	5	6	7	8	9	10	11–15	16–20	21–25	26–30	30+
Good terms	25	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	34	2	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	63	3	2	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
	123	10	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	168	11	2	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
Average terms	286	351	37	21	14	5	0	0	1	0	0	1	0	0	0	0	0	0	0
	11	352	34	22	16	3	1	1	1	0	0	0	0	0	0	0	0	0	0
	23	353	31	22	16	4	1	1	0	0	0	0	0	0	0	0	0	0	0
Poor terms	253	734	180	92	46	27	10	4	1	0	3	1	0	0	0	0	0	0	0
	388	735	192	92	43	23	13	7	4	0	1	0	1	0	0	0	0	0	0
	389	736	302	116	48	15	18	18	9	4	3	1	0	0	0	0	0	0	0

TABLE 4

Comparison of average rank for top 25 and bottom 25 terms for DV, EK, and S/N measures (two document collections)

	CRAN 425			MED 450		
	DV	*EK*	*S/N*	*DV*	*EK*	*S/N*
Top 25 *DV*	12.5	53.5	97.8	12.5	132.0	221.0
Worst 25 *DV*	2638.5	492.0	835.0	4713.5	712.0	2803.0
Top 25 *S/N*	211.0	16.5	12.5	128.6	16.6	12.5
Worst 25 *S/N*	704.0	2353.0	2638.5	3709.0	3025.0	4713.5
Top 25 *EK*	147.0	12.5	14.3	483.0	12.5	23.0
Worst 25 *EK*	653.8	2638.5	2625.0	1870.0	4713.5	4694.0

which gives the average ranks of the top 25 and bottom 25 terms ranked acco to the *DV*, *EK* and *S/N* measures for two document collections in aerodyn (CRAN 425) and medicine (MED 450). The average rank for the top 25 is of 12.5. For the bottom 25, the average is 2638.5 and 4713.5 for the CRAN and collections which contain a total of 2,651 and 4,726 terms in all. The sign calculations produce approximately equivalent average ranks for metho are reasonably similar; for methods that are not comparable, the 25 bes according to one ranking system may, however, be ranked in the middle, or the bottom of the list according to some other system.

The data of Table 4 may be summarized in the following way:

(a) Terms with high *DV* values have fair to average *EK* values and ave weights; terms with low *DV* values are mediocre according to *EK* poor in *S/N*.

(b) Terms with good *S/N* values have good *EK* values and fair to av weights; the poor *S/N* terms are also poor according to *EK* and f in *DV* weight.

(c) Good *EK* terms also have good *S/N* values and fair to average poor *EK* terms are also poor *S/N* terms and quite poor discrimi

Thus, there appears to be almost perfect agreement between the effect o noise and the variance based *EK* measures. The differences between the tion values (*DV*) and the other two procedures (*EK* and *S/N*) are more but even there the high discriminators have at least average value acc and *S/N*, and poor discriminators are also quite poor in *EK* and *S/N*

A more detailed comparison between the *S/N* and *DV* methods in Table 5. In each case, the frequency distributions of some typical and poor *S/N* terms are given in the upper half of the table; the s presented for the *DV* terms in the bottom half of the table. The ter beginning of the table is the best *S/N* term in the collection unde number 195), and it occurs once in one document, twice i

TABLE 5

Frequency distributions of sample terms exhibiting good, average, or poor S/N and DV characteristics (CRAN 1400 collection—736 distinct term classes)

Characterization	Term number	S/N rank	DV rank	F	B	1	2	3	4	5	6	7	8	9	10	11-15	16-20	21-25	26-30	30+
						Number of documents in which term appears with f_i^k of														
Good S/N	195	1	151	20	3	1	1	0	0	0	0	0	0	0	0	0	1	0	0	0
	598	2	91	33	6	2	2	0	0	0	0	0	1	0	0	0	1	0	0	0
	639	3	383	9	2	1	0	0	0	0	0	0	1	0	0	0	0	0	0	0
	461	10	197	42	13	4	4	0	3	0	1	0	0	0	0	1	0	0	0	0
	390	11	1	416	97	27	18	9	7	7	5	1	3	6	3	5	3	0	0	0
Average S/N	507	351	147	277	176	123	33	10	3	2	2	2	0	0	1	0	0	0	0	0
	159	352	153	87	55	30	18	7	0	0	0	0	0	0	0	0	0	0	0	0
	88	353	104	128	83	57	14	7	3	2	0	0	0	0	0	0	0	0	0	0
Poor S/N	521	734	252	164	138	116	18	4	0	0	0	0	0	0	0	0	0	0	0	0
	54	735	247	143	122	105	13	4	0	0	0	0	0	0	0	0	0	0	0	0
	656	736	409	14	14	14	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Good DV	390	11	1	416	97	27	18	9	7	7	5	1	3	6	3	5	3	0	0	0
	281	36	2	572	189	82	36	20	22	9	4	4	1	2	2	4	1	0	1	1
	69	12	3	185	52	22	12	2	3	3	2	1	2	0	1	2	2	0	0	0
	197	113	10	243	100	39	28	15	7	2	3	3	2	0	1	0	0	0	0	0
	238	105	11	261	107	47	23	14	7	8	3	2	1	2	0	0	0	0	0	0
Average DV	371	644	351	30	25	20	5	0	0	0	0	0	0	0	0	0	0	0	0	0
	397	604	352	14	12	10	2	0	0	0	0	0	0	0	0	0	0	0	0	0
	321	91	353	17	8	3	3	1	0	1	0	0	0	0	0	0	0	0	0	0
[illegible] DV	276	44	734	1560	420	110	77	55	57	38	24	16	13	7	7	13	3	0	0	0
	[illegible]	[illegible]	735	2359	527	113	114	58	51	39	44	16	26	14	10	28	10	3	0	1
						[illegible]	178	110	79	46	38	18	10	3	2	3	0	0	0	0

between 16 and 20 times in a third document. At the bottom of the table the worst discriminator with DV rank 736 (term number 389) is a high-frequency term which occurs once in 235 different documents, twice in 173 other documents, three times in 110 more, four times in 79 others, and so on down to the three last documents in which its occurrence frequency is between 11 and 15. Out of the 1,400 documents used in the collection examined in Table 5, term 389 is in fact assigned to over half the items (719 documents).

From the data of Table 5 it is clear that the best S/N terms have very low document frequencies and not very high discrimination values for the most part. This confirms the previously made comment that the S/N and EK formulas favor high concentration. The average S/N terms exhibit a medium document frequency and a total collection frequency which is about fifty percent higher than the document frequency. Their frequency distributions are characterized by an occurrence frequency of 1 in a very large proportion of the documents to which they are assigned. This last feature is accentuated even more in the poor S/N terms—these terms occur exclusively with very low term frequencies, and the distribution is very flat.

The characterization of the S/N terms contained in the upper half of Table 5 makes it appear that the S/N classification is one based on specificity alone, and that it is not well correlated with the frequency characteristics. In a retrieval situation, the good S/N terms may be as ineffective (because they occur so rarely) as the poor S/N terms that occur so often with a frequency equal to 1.

Consider now the DV characteristics shown at the bottom of Table 5. The best DV terms have average document frequency, and a collection frequency at least two to three times higher than the document frequency. Furthermore, they exhibit skewed frequency distributions in that the frequencies of occurrence vary from very low in some documents to quite high in some others.

The average DV terms have low document frequencies, and total collection frequencies approximately equal to the document frequencies. For practical purposes, the average discriminators are terms that occur with a term frequency of 1 in relatively few documents in a collection.

The poor discriminators, finally, have high document frequency, and collection frequencies two or three times the size of the document frequency. The number of documents in which these terms occur with low frequency is very large, which of course accounts for their low discrimination values. Whereas no clear correlation was found to exist between the S/N ratings and the document or collection frequencies of the corresponding terms, a direct relation appears to exist for the discrimination value rankings. As the discrimination values decrease from good to average to poor, the document and collection frequencies of the terms go from average, to low, and finally to quite high. This correspondence is used as a basis for a theory of indexing in the last section of this study.

In summary, a study of the frequency distributions of the terms ranked according to a number of different measures of term significance reveals the following characteristics:

(a) When the terms are ranked in decreasing order of collection frequency F^k, or document frequency B^k, the best terms are those with universal occurrence characteristics: such terms may help in producing high recall output, but the retrieval results will certainly not be sufficiently precise for most purposes.

(b) A ranking in inverse collection or document frequency ($1/F$ or $1/B$) puts at the top of the list terms with total occurrence frequencies equal to 1: such terms are not useful in obtaining effective retrieval output because of their excessive rarity.

(c) The variance-based (EK) and signal-noise (S/N) measures have identical occurrence characteristics, favoring completely concentrated terms in both cases; while those terms may be usable to generate high precision output, they appear to be too specific and too rare to help an average user in searching an average collection.

(d) The discrimination value (DV) ranking appears to reflect those term characteristics normally thought to be important in retrieval — the best terms being those with skewed frequency distributions that occur neither too frequently nor too rarely: the least attractive terms from the discrimination point of view are terms occurring everywhere that are not capable of distinguishing the items from each other.

(e) The information value (IV) process must be based on a large number of user-system interactions; reliable frequency distribution characteristics remain to be generated in this case.

A final standard of comparison for the significance measures relates to the computational complexity. Let

t be the total number of distinct terms assigned to the documents,

n be the total number of documents,

K be the average length of the document vectors (that is, the average number of nonzero terms),

and

K' be the average document frequency of a term (that is, the average number of documents to which a term is assigned).

In increasing order of difficulty, the following computational requirements become necessary: for the weighting system based on collection or document frequencies (formulas (4) and (5)), K' additions are needed per term; for t terms, this produces $K't$ additions.

To compute the EK value in accordance with formula (11) the total requirements are

K' additions to compute F^k,

K' multiplications for the $(f_i^k)^2$ terms,

K' additions for $\sum_{i=1}^{n} (f_i^k)^2$,

1 division for n/F^k,

1 multiplication to complete the first term in (11),

1 subtraction.

The total is $2K' + 1$ additions or subtractions, and $K' + 2$ multiplications or divisions. For t terms, this produces $(2K' + 1)t$ additions and $(K' + 2)t$ multiplications. The last term represents the increment over and above the simple frequency counts of expressions (4) and (5).

The signal-noise calculations are more expensive to perform than the EK values. Consider first the noise N^k (formula (6)); the requirements are

	K'	additions for F^k,
	$2K'$	divisions,
	K'	logarithms,
	K'	multiplications,
and	K'	additions to compute the final sum.

In addition, the computation of the signal S^k (formula (7)) adds K' logarithms and 1 subtraction. The total requirements are then equal to $2K' + 1$ additions or subtractions, $3K'$ multiplications or divisions, and $2K'$ logarithms. For t terms, this produces $(2K' + 1)t$ additions, $3K't$ multiplications, and $2K't$ logarithms. If the figure of merit FM of formula (8) is used, t multiplications and t divisions must be added.

Consider finally the computations needed for the discrimination value. The centroid C of the document space, defined as the average document, requires n additions for each of t terms, or a total of $t \cdot n$ additions, plus optionally t divisions.

The space compactness function Q_k (formula (13)) may be defined as

$$Q = \sum_{j=1}^{n} \cos(C, D_j) = \sum_{j=1}^{n} \left\{ \frac{\sum_{i=1}^{t} (c_i \cdot d_{ij})}{\sqrt{\sum_{i=1}^{t} (c_i)^2 \sum_{i=1}^{t} (d_{ij})^2}} \right\}, \tag{18}$$

where the similarity function s of expression (13) is replaced by the cosine function. The outside summation is assumed to encompass all documents. The following operations appear to be needed:

numerator:	K	multiplications and K additions,
denominator:	t	multiplications and t additions for the sum over (c_i^2),
	K	multiplications and K additions for the sum over (d_i^2),
	1	multiplication and 1 square root,
ratio:	1	division.

All operations involving the document terms d_i must be repeated for all n documents, and the final sum of n terms must be obtained. This produces the following totals for the computations of Q:

$(2K + 1)n + t$	multiplications,
$(2K + 1)n + t$	additions,
n	square roots,
n	divisions.

In addition to computing the space density Q, it is also necessary to generate Q_k, the density with term k removed, for all terms k. The basic definition is

$$Q_k = \sum_{j=1}^{n} \cos_k (C, D_j)$$

$$(19) \qquad = \sum_{j=1}^{n} \left\{ \frac{\sum_{i \neq k, i=1}^{t} (c_i \cdot d_{ij})}{\sqrt{\sum_{i \neq k, i=1}^{t} (c_i)^2 \cdot \sum_{i \neq k, i=1}^{t} (d_{ij})^2}} \right\}$$

$$= \sum_{j=1}^{n} \left\{ \frac{\sum_{i=1}^{t} (c_i \cdot d_{ij}) - c_k d_{kj}}{\sqrt{\left(\sum_{i=1}^{t} (c_i)^2 - c_k^2\right)\left(\sum_{i=1}^{t} (d_{ij})^2 - d_{kj}^2\right)}} \right\}$$

The formula of expression (19) makes it clear that if the possibility existed of storing the sums inside the braces which are already contained in (18), the t computations of Q_k would add essentially a factor of t to the number of operations required. There are, however, n sums for $\sum c_i d_{ij}$, and n for $\sum d_{ij}^2$, and the storage space required for this purpose may not be available. The single sum for the centroid $\sum c_i^2$ may, however, be saved in all cases.

Using the same calculations as before, the following operations are necessary for a complete computation of Q_k:

numerator:	$(K + 1)n$ multiplications,
	$(K + 1)n$ additions or subtraction,
denominator:	1 multiplication and 1 addition for the sum over c,
	$(K + 1)n$ multiplications and
	$(K + 1)n$ additions or subtractions,
	n multiplications,
	n square roots,
ratio:	n divisions.

The work must be repeated t times for all t terms, and t final subtractions are necessary to compute $(Q_k - Q)$ for all terms. The totals are then as follows:

$(2Kn + 4n + 1)t$	multiplications or divisions,
$(2Kn + n + 2)t$	additions or subtractions,
nt	square roots.

The final operational complexity for t computations of $Q_k - Q$ is then

$(2Kn + 4n + 2)t + 2Kn + 2n$	multiplications or divisions,
$(2Kn + n + 3)t + 2Kn + n$	additions or subtractions,
and $(n + 1)t$	square roots.

A summarization of the complexity of the significance computations is given in Table 6. Since the discrimination value measure is dependent on the collection

TABLE 6
Computational complexity of significance computations

Significance measure	Computational requirements		Overall order (multiplications)
F or B	$K't$	additions	—
EK	$(2K' + 1)t$	additions	$o(K't)$
	$(K' + 2)t$	multiplications	
S/N	$(2K' + 1)t$	additions	
	$3K't$	multiplications	$o(3K't)$
	$2K't$	logarithms	
DV	$(2Kn + 4n + 2)t + 2Kn + 2n$	multiplications	
	$(2Kn + n + 3)t + 2Kn + n$	additions	$o(2Knt)$
	$(n + 1)t$	square roots	

size, the calculations become automatically much more demanding than those required for the other measures.

5. Experimental results. The term significance measures previously introduced can be used in various ways to enhance retrieval performance in an information processing environment. In particular, by choosing a threshold in the significance values, terms of low or inadequate significance can be removed from the indexing vocabulary to produce a better or more effective vocabulary. The choice of a variety of thresholds leads to the so-called CUT experiments described in this section. As suggested earlier, the significance values can also be applied as an element in computing weighting factors to be assigned to the terms characterizing each document. Thus, the standard term frequency factor f_i^k of term k in document i might be refined by multiplication with one of the collection-dependent significance measures such as the discrimination value, or the signal-noise ratio. The combination of document-related and collection-related measures is designated a: MULT in the experimental output.

Except where otherwise noted, the experimental results are based on the use o three collections of about 450 documents each in aerodynamics, biomedicine and world affairs, respectively, denoted as CRAN, MED, and Time; twenty-fou queries are used with each collection. While different subject areas are covered i each case, the relevance properties are identical for the three collections; i particular, the probability that a given document is relevant to a query is the sam throughout the test base. The basic collection statistics are shown in Table 7.

The experiments are based on standard word stem indexing in which wor stems are automatically extracted from document abstracts to serve as index terr

TABLE 7
Basic collection statistics for three test collections

Collection statistics	CRAN 424	MED 450	Time 425
Subject area	aerodynamics	biomedicine	world affairs
Number of documents	424	450	425
Average document length in words	200	210	570
Number of queries	24	24	24
Relevance count (average number of relevant documents per query)	8.7	9.2	8.7
Generality (relevance count divided by collection size)	0.02	0.02	0.02

[12, Chap. 3]. The basic indexing statistics are shown for the three collections in Table 8. It may be seen that the total number of distinct terms (word stems) used to index the three collections increases from CRAN to MED, and from MED to Time. In the last case, the indexing vocabulary was artificially limited in size by removing terms with a total collection frequency F^k equal to 1 (but not those whose document frequency B^k was equal to 1, with F^k larger than 1). The average term frequency is approximately equal for CRAN and Time; but for the MED collection it is much lower, indicating that a large number of low frequency terms are used to represent the documents of that collection.

TABLE 8
Basic indexing statistics

Indexing statistics	CRAN 424	MED 450	Time 425
Number of distinct terms (word stems)	2,651	4,726	7,569
Total number of term occurrences	35,353	29,193	112,136
Average term frequency	14.8	6.2	13.3
Average number of terms per document	83.4	64.8	263.8
Compression percentage of documents (indexing length to word length)	40%	30%	46%

Various experimental results are examined in the remainder of this section.

A. *Binary versus term frequency indexing.* The first question that might be raised concerns the usefulness of the term frequency weighting compared with the standard binary weighting. The following two questions may be considered in particular:

(a) Are the term frequency weights f_i^k generally useful to enhance recall beyo the performance obtainable with ordinary binary weights b_i^k?
(b) To what extent can the upweighting of very high frequency terms with lo discriminatory power implicit in the term frequency weighting be mitigate by using a factor in inverse document frequency order in addition to th term frequency weights?

Recall-precision tables are included for the three experimental collections i Table 9. In each case, precision values are given at ten recall points spaced in step of 0.1, averaged over the 24 user queries that are utilized with each collectio

TABLE 9
Comparison of binary and term frequency weighting with and without inverse document frequency normalization

	R	Binary weights b_i^k	Term frequency weights f_i^k	Binary with IDF weights $b_i^k \cdot (IDF)_k$	Term frequency with IDF $f_i^k \cdot (IDF)_k$
CRAN	.1	.7165	.6844	.7502	.7573
	.2	.5419	.5303	.6692	.6241
	.3	.4581	.4689	.5336	.5348
	.4	.3673	.3482	.4146	.4457
	.5	.3231	.3134	.3475	.3935
	.6	.2664	.2556	.2946	.3182
	.7	.2283	.1989	.2431	.2521
	.8	.2082	.1631	.1923	.1953
	.9	.1538	.1265	.1409	.1388
	1.0	.1439	.1176	.1328	.1277
MED	.1	.7958	.7891	.7770	.8459
	.2	.6912	.6750	.7069	.7557
	.3	.5772	.5481	.6037	.6584
	.4	.5339	.4807	.5453	.5442
	.5	.4880	.4384	.5315	.4873
	.6	.3777	.3721	.4179	.4254
	.7	.3350	.3357	.3897	.3833
	.8	.2421	.2195	.2795	.2620
	.9	.1916	.1768	.2080	.2126
	1.0	.1391	.1230	.1490	.1469
Time	.1	.8257	.7496	.8085	.8536
	.2	.7555	.7071	.7741	.7901
	.3	.6754	.6710	.7114	.7568
	.4	.6224	.6452	.6328	.7305
	.5	.5708	.6351	.6218	.6783
	.6	.5299	.5866	.5673	.6243
	.7	.4618	.5413	.5124	.5823
	.8	.4087	.5004	.4384	.5643
	.9	.2959	.3865	.3374	.4426
	1.0	.2854	.3721	.3188	.4170

Four weighting procedures are used to produce the output of Table 9, including binary term weights b_i^k, term frequency weights f_i^k, and binary as well as term frequency weights multiplied by an inverse document frequency factor, designated $(IDF)_k$ in Table 9. A weighting system such as $(F_i^k) \cdot (IDF)_k$ may be expected to produce high recall (because of the f_i^k factor) as well as high precision (because of the IDF factor).

To represent the inverse document frequency, an integral weighting function IDF is used, where

$$(IDF)_k = f(n) - f(B^k) + 1. \tag{20}$$

n is the number of documents in the collection, and $f(x) = \lceil \log_2(x) \rceil$. Obviously, expression (20) takes on small values for terms with large B^k, and large values when B^k is small (see [1]).

No simple answer can be given to question (a) above concerning the superiority of binary or term frequency weighting. The curly line in the b_i^k and f_i^k columns of Table 9 designates the better precision values in each case. It may be seen that for the CRAN and MED collections, the binary weights are normally superior, whereas for the Time collection the term frequency weighting is preferable. However, the differences in performance are large only for the Time collection. This may be ascertained by consulting column 1 of Table 10 which contains statistical significance test results for certain pairs of weighting methods.

TABLE 10
Statistical significance output for the results of Table 9

	A. Binary weights b_i^k vs B. Term freq weights f_i^k		A. Binary with IDF vs B. Term freq with IDF		A. Term freq f_i^k vs B. Term freq with IDF $(f_i^k \cdot IDF)$	
CRAN	t-test (B > A)	.9549	t-test (B > A)	.1580	t-test (B > A)	.0000
	Wilcoxon	.1701	Wilcoxon	.0146	Wilcoxon	.0105
MED	t-test (A > B)	.0626	t-test (B > A)	.3126	t-test (B > A)	.0000
	Wilcoxon	.4032	Wilcoxon	.4412	Wilcoxon	.0000
Time	t-test (B > A)	.0000	t-test (B > A)	.0000	t-test (B > A)	.0000
	Wilcoxon	.0000	Wilcoxon	.0001	Wilcoxon	.0000

Table 10 contains t-test and Wilcoxon signed rank test values, giving in each case the probability that the output results for the two test runs could have been generated from the same distribution of values. Small probabilities—for example, those less than 0.05—indicate that the answer to this question is negative and that the test results are significantly different [24]. It may be seen in Table 10 that only for the Time collection is there a significant difference between binary and tem frequency weighting, with the latter being substantially better than the forme (B > A).

When the use of the inverse document frequency factor is considered, as shown in the last two columns of Table 9, it may be seen that substantial improvemens in performance are produced. That is, term weights equal to $(b_i^k \cdot IDF_k)$ are generally superior to (b_i^k) alone; the same is true of $(f_i^k \cdot IDF)_k$ over (f_i^k) alone. The difference between the last two systems are statistically fully significant, as indicated in column 3 of Table 10.

The best of the four frequency-based weighting systems is identified in Table 9 by a vertical bar. It may be seen that the bar is generally concentrated in the las column. The following overall conclusions appear to be warranted:

(a) whether term-frequency weighting (f_i^k) is useful, compared with standad binary weights (b_i^k) depends on the collection and query characteristics;
(b) when inverse document frequency weighting (IDF) is used, $(b_i^k \cdot IDF_k)$ is generally superior to b_i^k alone, and $(f_i^k \cdot IDF_k)$ is always superior to f_i^k;
(c) the best performance is obtained with a combined term frequency weightin for recall, with inverse document frequency for precision $(f_i^k \cdot IDF_k)$; thi system prefers terms with high individual term frequencies and low overal document frequencies.

The frequency-based weights are compared with other weighting systems in th remainder of this section.

B. *Term deletion experiments.* All existing indexing theories make specia provisions for the removal of certain high-frequency terms that are believed not t be useful for content identification. Thus, "stop lists" or "negative dictionaries are used to delete a number of common words, normally including preposition conjunctions, articles, auxiliary verbs, etc., before some of the remaining terms ma be chosen for content identification. The number of common function wor included in a standard stop list may range from 50 to about 200, depending on th system in use.

Since the significance measures described previously can be used to assign each term a value reflecting its importance for content analysis purposes, one m inquire whether savings are possible by reducing the indexing vocabulary to sor optimum size. In particular, following the elimination of the common wor included on the stop list, the remaining terms might be arranged in decreasi order of their term weights—for example, in decreasing discrimination order—a terms whose value falls below some given threshold might be eliminated.

The characteristics of low-valued terms vary with the particular indexi strategy—in general, they may be high frequency terms that occur everywh (that is, they are assigned to all items in a collection), or they may, on the contra be very low-frequency terms that occur only once or with low frequency. In eit case, these terms use up considerable storage space, and they may contrib little to the retrieval effectiveness.

A typical strategy used experimentally with a collection of 1,033 docum abstracts in biomedicine is shown in Fig. 8 (from [25]). In this system about

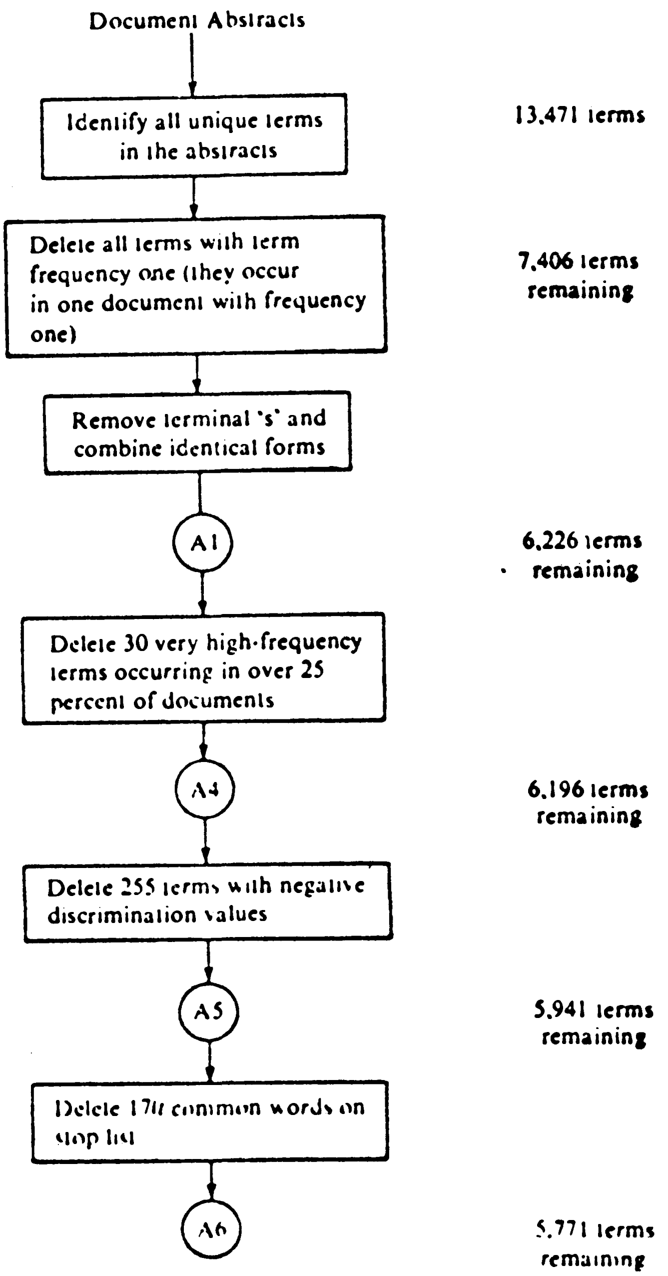
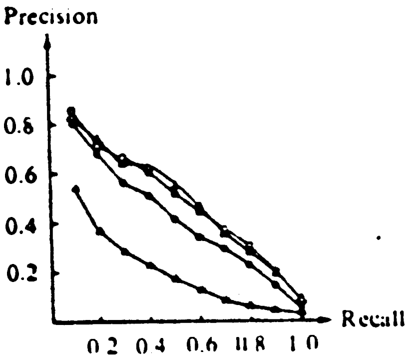


FIG. 8. *Typical term deletion algorithm (adapted from [25]).*

percent of the unique words contained in the original document abstracts are used for indexing purposes, the largest amount of deletion being obtained by eliminating terms of frequency one. Such terms do not provide much matching power between documents and queries—in fact, when they occur in a query, they may help in the retrieval of one document at most. Additional deletions are carried out by removing terms with a large document frequency, standard common words, terms with negative discrimination values, and terms that differ from existing ones only by addition of a terminal 's'.

Recall-precision results averaged for 1,033 document abstracts and 35 user queries are shown for the system in Fig. 9. A recall-precision graph such as the one in Fig. 9 is simply a graphic representation of the standard recall-precision tables in which adjacent precision values are joined by a line. The curve closest to the upper-right-hand corner of the graph (where recall and precision are highest) reflects the best performance. It may be seen in Fig. 9 that the deletion of frequency-one terms and of terms with large document frequencies produces substantial increases in the average recall and precision values.

▲——▲ A1 – Freq. one terms deleted.
●——● A4 – Freq. one and high freq. terms deleted.
△——△ A5—Freq. one, high freq., and nondiscriminators deleted.
○——○ A6—Freq. one, high freq., nondiscriminators, and standard common words deleted.



R	Precision			
	▲—▲	●—●	△—△	○—○
0.1	.5508	.8264	.8304	.8415
0.3	.2994	.5779	.6689	.6716
0.5	.1750	.4153	.5332	.5246
0.7	.0913	.3026	.3717	.3768
0.9	.0421	.1501	.2022	.2023

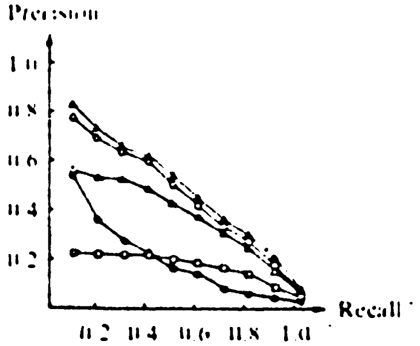
FIG. 9 *Performance of term deletion algorithm of Fig. 8; averages over 1033 documents and 35 queries (adapted from [25]).*

Additional reductions in the indexing vocabulary may be effected by further deletion of terms in increasing term value order. Thus the 5,941 terms constituting the A5 word list of Fig. 8 might be reduced to only 1,000 terms by deleting the 4,941 terms that exhibit the next lowest discrimination values.

The recall-precision output of Fig. 10 reflects the retrieval performance for the previously used collection of 1,033 items in biomedicine, again averaged over 35 search requests. It is seen that only a few percentage points are lost when the indexing vocabulary is reduced from the original 13,400 distinct words occurring in the document abstracts to the 1,000 terms exhibiting the best discrimination values. As additional terms are deleted in increasing discrimination value order, it becomes apparent that important content words (good discriminators) are affected because the performance drops drastically when the indexing vocabulary is reduced to 500 terms, and it is very poor indeed when the best 250 terms only are utilized.

The results of Figs. 9 and 10 give no clue concerning the optimum size of the indexing vocabulary to be used for any given collection. To study this question a

●—● 6226 Terms (A1)
▲—▲ 5941 Terms (A6)
○—○ 1000 Terms
△—△ 500 Terms
□—□ 250 Terms



	Precision				
R	●—●	▲—▲	○—○	△—△	□—□
0.1	.5508	.8308	.7919	.5531	.2145
0.3	.2904	.6689	.6670	.5307	.2108
0.5	.1750	.5332	.5126	.4286	.1979
0.7	.1013	.3717	.3661	.3292	.1793
0.9	.0421	.2022	.2017	.1813	.0925

FIG. 10 *Reduction of terms by deletion of poor discriminators; averages over 1033 documents and 35 queries (adapted from [25]).*

variety of different deletion thresholds are used with the three test collections previously introduced. In all cases, standard binary term weights (b_i^k) are utilized, and deletion occurs in inverse document frequency order that is, terms whose document frequency is greater than a given threshold are deleted.

The term deletion statistics are given in Table 11, and the corresponding recall-precision results are shown in Table 12 [26]. An asterisk in Tables 11 and 12 identifies the three runs for which the deletion percentage is approximately equal—about 11 percent of the total term occurrences. The output of Table 12 shows that no unified policy appears to be derivable from the test results. Indeed, for the CRAN collection, the best policy consists in not deleting any terms at all, whereas the best results for MED and Time are obtained for deletions of terms with document frequencies $B^k \geq 16$ and $B^k \geq 104$, respectively, corresponding to the elimination of about ten percent of total term occurrences. Since such a relatively small deletion percentage does not lead to substantial losses in performance for any collection, and may in fact produce considerable improvements, the ten percent deletion percentage may be productive in all environments.

It may be useful, as a final exercise, to determine whether a clear-cut policy is available for choosing among various significance rankings for term deletion purposes. In particular, the discrimination value rankings can be compared with the inverse document frequency rankings previously examined. The output of Table 13 shows two of the most effective term deletion runs using both inverse document frequency (*IDF*) rankings, and discrimination order (*DISC*) rankings. In each case, term frequency weights are used for indexing purposes (rather than binary weights as in Table 12). The deletion thresholds for removing terms with high document frequency are $B^k \geq 129$, 19, and 104 for CRAN, MED, and Time, respectively. This removes 0.50, 3.70 and 0.33 percent of the terms with highest document frequency, accounting for 11.80, 9.71, and 11.1 percent of the total

Table 11

Term deletion statistics (deletion in IDF order; standard binary term weighting)

Collection	Number of distinct terms	Number of term occurrences	Number of terms deleted	Document frequency threshold	Percentage of term occurrences deleted	Average collection frequency of terms deleted	Average document frequency of terms deleted
CRAN	2651	35,353	13 (0.49%)	129	11.8*	320.8	158.2
			71 (2.67%)	60	35.3	175	99
			104 (3.92%)	49	44.8	152	84.9
			128 (4.82%)	41	49.3	136	77.3
MED	4726	29,193	133 (2.81%)	23	8.38	70.7	39.6
			175 (3.7%)	19	9.71	62.2	35
			228 (4.82%)	16	10.94*	53.8	30.8
Time	7569	112,136	45 (0.6%)	104	11.1*	276.7	141.5
			207 (2.73%)	56	28.6	155	88.7
			255 (3.36%)	51	31.9	140.2	82.2
			389 (5.13%)	41	39.5	114	69.3

* Same percentage of deleted terms.

TABLE 12

Term deletion results (deletion in IDF order: binary term weighting)

	Recall	Standard binary b^2	IDF CUT $w \geq$ [illegible]	IDF CUT $w \geq 60$	IDF CUT $w \geq$ [illegible]	IDF CUT $w \geq 41$
CRAN	.1	.7165	6811	7516	.7169	6821
	.2	5419	.5545	6276	5893	5369
	.3	4581	4832	4484	4446	4222
	.4	3673	3719	3545	3464	3249
	.5	3231	3046	2729	2835	2725
	.6	2664	2536	2334	2350	2349
	.7	2283	2021	2039	1804	.1845
	.8	2082	1823	1782	1194	.1206
	.9	1538	1335	1351	1056	.1128
	1.0	1439	1215	1315	1056	.1128
	Recall	Standard binary b^2	IDF CUT $w \geq$ [illegible]	IDF CUT $w \geq$ [illegible]	IDF CUT $w \geq$ [illegible]	
MED	.1	.7958	.7778	7872	.7441	
	.2	6912	6954	6692	6736	
	.3	5772	6253	6197	.5739	
	.4	5339	5871	5948	.5423	
	.5	4880	.5228	5299	4801	
	.6	3777	4542	4628	3990	
	.7	3350	.4361	4377	.3833	
	.8	.2421	2862	.3084	2587	
	.9	.1916	.2107	.2252	1971	
	1.0	.1391	.1358	.1385	.1245	
	Recall	Standard binary b^2	IDF CUT $w \geq$ [illegible]	IDF CUT $w \geq$ [illegible]	IDF CUT $w \geq$ [illegible]	IDF CUT $w \geq 41$
Time	.1	8257	8306	7614	.7445	.6642
	.2	7555	7690	7368	.7326	6634
	.3	6754	7084	6529	.6559	.6157
	.4	.6224	6164	.5895	.5901	.5387
	.5	.5708	.5955	.5258	.5373	.4701
	.6	.5299	.5529	.4991	.5060	.4406
	.7	.4618	.4737	.4279	.4294	.3970
	.8	.4087	.4158	.3643	.3620	.3190
	.9	.2950	.3025	.2909	.2837	.2446
	1.0	.2854	.2928	.2860	.2685	.2404

TABLE 13
Recall-precision results for two term deletion methods using three test collections

	R	Standard binary weights	Standard term frequency weights	Term frequency weights: IDF CUT	Term frequency weights: DISC CUT	IDF CUT vs Standard term frequency	DISC CUT vs Standard term frequency
CRAN	.1	.7165	.6844	.6975	.6654		
	.2	.5419	.5303	.5945	.5733	*t*-test	t-test
	.3	.5481	.4689	.5097	.5142		
	.4	.3673	.3482	.4197	.4654	.0000	.2841
	.5	.3231	.3134	.3355	.3542		
	.6	.2664	.2556	.2938	.2923	Wilcoxon	Wilcoxon
	.7	.2283	.1989	.2326	.2341		
	.8	.2082	.1631	.1802	.1492	.0115	.6561
	.9	.1538	.1265	.1316	.1274		
	1.0	.1439	.1176	.1256	.1223		
MED	.1	.7958	.7891	.7999	.8691		
	.2	.6912	.6750	.7622	.8115	*t*-test	*t*-test
	.3	.5772	.5481	.6865	.6677		
	.4	.5339	.4807	.6083	.6136	.0001	.0000
	.5	.4880	.4384	.5603	.5798		
	.6	.3777	.3221	.4682	.4912	Wilcoxon	Wilcoxon
	.7	.3350	.3357	.4423	.4474		
	.8	.2421	.2195	.3139	.2988	.0001	.0000
	.9	.1916	.1768	.2452	.2325		
	1.0	.1391	.1230	.1524	.1499		
Time	.1	.8257	.7496	.8601	.7911		
	.2	.7555	.7071	.8268	.7485	t-test	t-test
	.3	.6754	.6710	.7503	.7362		
	.4	.6224	.6452	.7144	.7000	.0000	.0085
	.5	.5708	.6351	.6872	.6777		
	.6	.5299	.5866	.6168	.6350	Wilcoxon	Wilcoxon
	.7	.4618	.5413	.5645	.5907		
	.8	.4087	.5004	.5017	.5510	.0001	.0127
	.9	.2959	.3865	.4071	.4177		
	1.0	.2854	.3721	.3906	.4019		

term occurrences, respectively. For the *DISC CUT* runs, the threshold is so chosen that all terms with a negative discrimination value are removed. Following removal of the respective terms, the remaining terms are used with standard term frequency weighting.

The recall-precision results shown in Table 13 for the three test collections show that in general better average performance is obtained when the low-valued terms are deleted than with the full vocabulary. The best performance result is emphasized in Table 13 by a vertical bar. The last two columns of the Table contain statistical significance output. For each pair of processes listed, *t*-test and Wilcoxon signed rank test probabilities are given. It is seen that all term deletion results are significantly better than the standard term frequency word stem weighting, with the exception of the *DISC* CUT run used with the CRAN collection.

While the term deletion systems appear to produce improvements in retrieval performance, it is again impossible to decide on an optimal deletion system based on the results of Table 13. In fact, for some recall values, the discrimination deletion is superior to the inverse frequency deletion, and vice versa for other recall areas. The question of what constitutes a good indexing vocabulary therefore requires further study.

C. *Multiplication experiments.* It was seen earlier that the collection-dependent significance measures can be used as multiplicative (or additive) factors in combination with document-dependent frequency weights to generate term values for indexing purposes. Such a combined measure favors terms that exhibit high weights both in individual documents, and also in the collection as a whole. A number of multiplicative weighting systems are examined in this subsection.

Table 14 contains recall-precision tables for four multiplicative indexing procedures, including $f_i^k \cdot IDF_k$, $f_i^k \cdot DV_k$, $f_i^k \cdot S/N_k$, and $f_i^k \cdot EK_k$. The standard term frequency weighting, f_i^k, is also included to serve as control. The last two columns of Table 14 cover procedures in which the term deletion method of Table 13 is combined with the multiplicative process. These runs are denoted $f_i^k \cdot IDF_k$ (CUT and MULT), and $f_i^k \cdot DV_k$ (CUT and MULT) respectively, to indicate that low-valued terms are deleted prior to the weight calculations. More complicated combinations of methods can be implemented, such as deletion in discrimination value order followed by weighting in inverse document frequency order (*DV* CUT and *IDF* MULT). These have been considered elsewhere [26].

The output of Table 14 makes it plain that the *S/N* and *EK* weights do not operate as effectively, on the whole, as the *DV* and *IDF* weightings. Furthermore, the choice among the last two procedures is not clear-cut. For CRAN and Time the inverse document frequency procedures are slightly preferable, whereas for MED, the discrimination value weighting is best. This last result is not surprising, if one remembers (from Table 8) that the MED collection contains mostly low frequency terms, so that nothing is gained by deemphasizing the high frequency components.

Of the methods included in Table 14, the best ones are those which combine deletion of low-valued terms with multiplication of frequency and significance weights. For CRAN and Time, the *IDF* CUT and MULT is preferred, whereas for the MED collection, the best results are obtained with *DV* CUT and MULT.

Statistical significance figures for the output of Table 14 are shown in Table 15. It is seen that the differences between the multiplicative *DV* and *IDF* methods and the standard term frequency weighting are statistically significant for all three collections, the improvement in average precision for the ten recall points ranging from 7 percent to 14 percent. For the CUT and MULT methods, the differences are significant for all but the *DV* CUT and MULT using the CRAN collection. The average improvement for the CUT and MULT methods over the standard term frequency weights is even larger, ranging from 8 percent to 23 percent.

Table 14

Recall-precision results for multiplication experiments

	R	Standard term frequency (TF) weights f_i^j	TF weights with IDF $f_i^j \cdot IDF_j$	TF weights with DV $f_i^j \cdot DV_j$	TF weights with S/N $f_i^j \cdot S/N_j$	TF weights with EK $f_i^j \cdot EK_j$	TF weights with IDF CUT + MULT $f_i^j \cdot IDF_j$	TF weights with DV CUT + MULT $f_i^j \cdot DV_j$
CRAN	.1	.6844	.7573	.6822	.6767	.6560	.7704	.6456
	.2	.5303	.6241	.6259	.5574	.5764	.6793	.5708
	.3	.4689	.5348	.5446	.5131	.5231	.5574	.5134
	.4	.3482	.4457	.4166	.4013	.4376	.4768	.4669
	.5	.3134	.3935	.3641	.3539	.3636	.3954	.3719
	.6	.2556	.3182	.3075	.2844	.2814	.3213	.3062
	.7	.1999	.2521	.2488	.2114	.2303	.2712	.2413
	.8	.1631	.1953	.1833	.1742	.1777	.2033	.1534
	.9	.1265	.1388	.1348	.1411	.1273	.1402	.1292
	1.0	.1176	.1277	.1279	.1335	.1197	.1306	.1240
MED	.1	.7891	.8459	.7995	.8042	.7270	.8275	.8322
	.2	.6750	.7557	.7255	.7562	.7138	.7548	.8113
	.3	.5481	.6584	.5949	.6369	.5647	.6764	.6671
	.4	.4807	.5442	.5066	.5566	.4876	.5968	.6230
	.5	.4384	.4873	.4530	.4969	.4252	.5457	.5834
	.6	.3721	.4254	.4053	.3911	.3668	.4789	.5119
	.7	.3357	.3833	.3715	.3391	.3128	.4336	.4690
	.8	.2195	.2622	.2460	.2618	.2209	.3066	.3087
	.9	.1768	.2123	.2033	.1981	.1756	.2390	.2401
	1.0	.1230	.1469	.1402	.1323	.1235	.1469	.1531
Time	.1	.7496	.8536	.8406	.7212	.7044	.8975	.8028
	.2	.7071	.7901	.7881	.7006	.6836	.8315	.7480
	.3	.6710	.7568	.7197	.6471	.6466	.7800	.7286
	.4	.6452	.7305	.6901	.6229	.6258	.7574	.6938
	.5	.6351	.6783	.6704	.6105	.5892	.7372	.6737
	.6	.5866	.6243	.6176	.5587	.5500	.6529	.6347
	.7	.5413	.5823	.5727	.5263	.4999	.5912	.5847
	.8	.5004	.5643	.5169	.4612	.4561	.5481	.5475
	.9	.3865	.4426	.4208	.3830	.3451	.4318	.4259
	1.0	.3721	.4170	.4053	.3593	.3186	.4118	.4085

TABLE 15
Statistical significance output for Table 14

	CRAN t-test	CRAN Wilcoxon	MED t-test	MED Wilcoxon	Time t-test	Time Wilcoxon
A. *TF* weights with *IDF*: $f_i^k \cdot IDF_k$	.0000	.0000	.0001	.0000	.0001	.0000
	A > B		A > B		A > B	
B. Standard *TF*: f_i^k	14%		12%		11%	
A. *TF* weight with *DV* $f_i^k \cdot DV_k$	.0000	.0000	.0000	.0000	.0000	.0000
	A > B		A > B		A > B	
B. Standard *TF*: f_i^k	11%		7%		8%	
A. *TF* with *IDF* CUT and MULT	.0000	.0000	.0008	.0001	.0001	.0000
	A > B		A > B		A > B	
B. Standard *TF*: f_i^k	19%		18%		15%	
A. *TF* with *DV* CUT and MULT	.1296	.4093	.0000	.0000	.0084	.0077
	A > B		A > B		A > B	
B. Standard *TF*: f_i^k			23%		8%	

To summarize, several methods based on the multiplication of standard term frequency weights by inverse document frequency and discrimination values have been found that appear to offer high performance standards. Among the methods which offer statistically significant improvements over the standard term weighting procedures for all processing environments, the following are the most promising:

(a) f_i^k standard weights with elimination of poor discriminators;

(b) $f_i^k \cdot IDF_k$ without elimination, or with elimination of poor discriminators or of terms with high document frequency;

(c) $f_i^k \cdot DV_k$ with elimination of poor discriminators or of high frequency terms.

D. *Information value experiments.* The experiments dealing with the use of information values are covered separately, because the methodology must necessarily be different in this case from that used earlier. In particular, since the generation of information values depends on a number of user-system interactions involving the processing of user queries against the available document collections, it is necessary to break the query set into two parts: a set of test queries must first be used for the generation and modification of term weights by means of interactive query processing; a new set of queries, not previously used, can then serve for evaluation purposes.

As explained earlier, the term (information) value generation process consists in increasing the weights of those terms which occur in queries and retrieved documents identified as relevant by the users; simultaneously, the weights are decreased when the terms cooccur in queries and retrieved documents identified as nonrelevant [27].

From an experimental viewpoint, two difficulties immediately arise. The first concerns the unavailability in many test environments of a sufficient number of user queries to carry out the interactive process. In the present instance, the information value test had to be abandoned for the MED collection because a sufficient number of user queries could not be found. The second problem is the relatively small number of cooccurring terms between documents and user queries, and thus the limited scope of the term value modifications. For the CRAN collection only about 20 terms in all were subjected to positive term modifications and only about 50 were modified negatively. The corresponding figures for Time are even smaller—about 10 positive modifications and about 30 negative ones. Obviously, stable information values cannot be obtained with such a small number of modification steps, with the result that the evaluation output may be considerably flawed.

For the CRAN collection, 131 test queries were used to generate the modified information values, while 59 test queries were available for this purpose with the Time collection. Twenty-four queries were used for the actual evaluation in each

TABLE 16
Information value experiments

	R.	Information value test 1	Information value test 2	Information value test 3	Term frequency and *IDF* $(f_2\ IDF_d)$
CRAN	.1	.6677	.6281	.6375	.7573
	.2	.6104	.5872	.5850	.6241
	.3	.5288	.4939	.4933	.5348
	.4	.4031	.4085	.4117	.4457
	.5	.3305	3254	.3146	.3935
	.6	.2918	.2496	.2529	.3182
	.7	.2020	.1980	.1962	.2521
	.8	.1409	.1377	.1384	.1953
	.9	.1038	.1901	.0891	.1388
	1.0	.0882	.0802	.0797	.1277
Time	.1	.8073	.8123	.8068	.8536
	.2	.7583	.7595	.7672	.7901
	.3	.7125	.7260	.7253	.7568
	.4	.6867	.6932	.6840	.7305
	.5	.6599	.6545	.6539	.6783
	.6	.6089	.6023	.5979	.6243
	.7	.5613	.5564	.5487	.5823
	.8	.5101	.5031	.5009	.5643
	.9	.3984	.4014	.4049	.4426
	1.0	.3757	.3698	.3692	.4170

case. For each test query, at most r relevant documents, and n nonrelevant documents retrieved above rank c were used to modify the information values. Three sets of values were tried for r, n, and c, as follows:

(a) test 1: $r = 2$, $n = 2$, $c = 5$,
(b) test 2: $r = 4$, $n = 4$, $c = 20$,
(c) test 3: $r = 8$, $n = 6$, $c = 40$.

The recall-precision results averaged over the 24 control queries are shown in Table 16. Also included in Table 16 is a term frequency-based control run $(f_i^k \cdot IDF_k)$.

It is clear from the results of Table 16 that the information value process does not lead to satisfactory output; in each case, the frequency-based weighting process is considerably superior. A final answer concerning the merits of the information values must await a larger test in a more realistic user environment.

6. A theory of indexing.

A. *The construction of effective indexing vocabularies.* The material presented up to now does not immediately lead to the generation of optimal indexing strategies valid in all environments. However, some generally useful conclusions are possible nevertheless:

(a) The only two significance measures leading to improvements in retrieval effectiveness are those based on inverse document frequencies (*IDF*) and on discrimination values (*DV*).
(b) The effectiveness of the significance measures for term deletion purposes (by removing low-valued terms from the indexing vocabulary) appears questionable, although a deletion percentage of about ten percent of total term occurrences does not lead to any serious performance deterioration.
(c) The main virtue of the significance measures is their function as collection-dependent weighting factors to be used in addition to the document-dependent term frequency values.

Even though the significance computations may not lead to optimal vocabularies by simple term deletion methods, one may ask whether good indexing vocabularies cannot be generated by *transforming* terms with low significance values, and thus high ranks, into new terms of better significance and lower rank. Specifically, a study of the formal characteristics of the terms arranged in order of significance may make it possible by suitable formal transformations to turn poor terms into better ones.

Consider first the terms in inverse document frequency (1/B or *IDF*) order, characterized by the frequency distributions of Table 3. The best terms are those with total frequency $F^k = B^k = 1$. While these terms exhibit low ranks, they are unlikely to provide optimal retrieval results because of their excessively low occurrence frequencies. Indeed, the virtue of the *IDF* significance measure for retrieval purposes appears to stem from its use as a combined weighting system with the standard term frequency values. A simple characterization of a useful retrieval term is thus difficult to generate directly from the *IDF* distributions of Table 3.

The situation is apparently less complicated when the terms are considered in order by discrimination value as represented in the lower half of Table 5. Obviously, the best terms have interesting frequency distributions, whereas the average and poor DV terms have either very low or very high occurrence frequencies. Furthermore, a direct correlation exists between discrimination value order and document frequency B^k. Indeed the distributions of Table 5 and the summarization of Table 17 indicate the following relations:

(a) The terms with the highest discrimination values (between 0.004 and 0.254 for the three test collections of Table 17) are those whose document frequency B^k is concentrated between 5 and 40 approximately for the test collections.[3]

(b) The terms with average discrimination ranks and discrimination values around zero are those with quite low document frequencies ranging from 1 to 5 for the test collections of Table 17.

(c) The terms with the lowest discrimination values (between −5.025 and 0 in Table 17) are characterized by the highest document frequencies ranging up to 270 for the collections of 450 documents.

The data of Table 17 also show that the class of high-frequency, negative discriminators is fairly small in each case. Because of their high individual document frequencies, these terms account, however, for a large proportion of total term occurrences. The class of low frequency terms with discrimination values near zero is normally large, while the number of good discriminators with medium document frequency is smaller in size. For the three sample collections of about 450 documents, the document frequency ranges applicable to the majority of the terms for the three classes of discrimination values are 1–5, 5–30, and 30–160, respectively.

If the discrimination value of a term furnishes an accurate picture of its value for indexing purposes, the situation may then be summarized, as shown schematically in Fig. 11. When the terms are arranged in increasing order according to their document frequencies in a collection, the first set of terms with very low document frequency B^k exhibits a discrimination value near zero. Next follow the terms with medium B^k and positive discrimination values; finally, the terms along the right-hand edge of Fig. 11 exhibit the poorest discrimination values and the highest document frequencies.

The document-frequency picture of Fig. 11 then suggests a model for the construction of good indexing vocabularies: the terms used for indexing purposes should as much as possible fall into the middle of the range of values represented in Fig. 11, by exhibiting low to medium document frequencies, and skewed term frequency distributions. This brings up two kinds of transformations that may be useful for improving existing indexing vocabularies [28]:

(a) a "right-to-left" transformation which takes high-frequency terms and breaks them apart into subsets, so that each subset exhibits a lower document frequency than the original; and

[3] The collection used to derive the data of Table 5 consisted of 1,400 documents, whereas only about 450 documents are included in each of the collections of Table 17. The document frequency values listed in the two tables are thus not compatible.

TABLE 17

Document frequency characteristics for terms in discrimination value order

	Term characteristics	Low document frequency terms Zero DV	Medium document frequency Positive DV	High document frequency terms Negative DV
CRAN 424	Discrimination value range	0–0.007	0.007–0.254	–2.936–0
	Number of terms in range	1990	587	74
	Document frequency range B^k	1-10	1–67	53-214
	Area of concentration of B^k	1–5	20–40	70–160
MED 450	Discrimination value range	0–0.008	0.008–0.138	–5.025–0
	Number of terms in range	3924	141	661
	Document frequency range B^k	1-26	1–28	14–138
	Area of concentration of B^k	1-3	5-20	20–70
Time 425	Discrimination value range	0–0.004	0.004–0.247	–1.862–0
	Number of terms in range	6468	725	406
	Document frequency range B^k	1-39	1–63	32-271
	Area of concentration B^k	1-3	5–30	32-140

(b) a "left-to-right" transformation which combines a number of low-frequency terms into supersets in such a way that each superset exhibits a higher document frequency than originally.

The right-to-left transformation which takes broad, high-frequency terms and renders them more specific should then be important as a precision-improving device, since the use of broad, nonspecific terms impairs the precision performance.

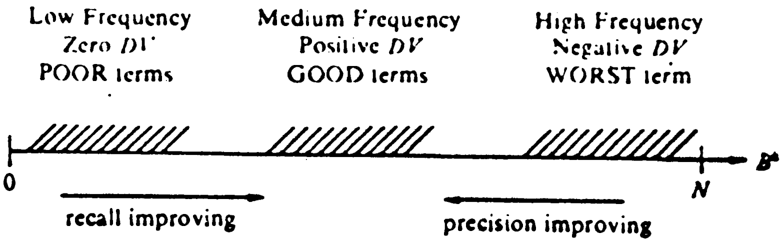


FIG. 11. *Term characterization in document frequency ranges*

Similarly, the left-to-right transformation should improve recall, because low-frequency specific terms are not helpful for recall purposes.

The proposed transformations are described and evaluated in the remainder of this section.

B. *Right-to-left phrase construction.* The right-to-left transformation takes high frequency terms and transforms them into units with lower frequency. The classical method for producing lower frequency terms from higher frequency components is to generate "phrases" consisting of several combined terms. For example, in a computer science collection, the terms "program" and "language" may be insufficiently specific, particularly when assigned to a large proportion of the documents in a collection. The phrase "programming language" is more specific and may, when assigned to the documents, lead to improved precision output. Unhappily, whereas a great deal is known about thesaurus construction (term grouping methods), the experiences obtained with phrase generation procedures have not been uniformly successful. Neither one of the two best-known phrase generation methods, involving either the use of syntactic analysis procedures for the formation of phrases or the use of statistical cooccurrence techniques, has been uniformly satisfactory in retrieval environments [24].

A new phrase generation system based on the term discrimination model is therefore proposed. Specifically, if the term characterization outlined in Fig. 11 is in fact an accurate representation of the indexing value of the terms it must be possible to improve the retrieval performance by breaking up terms with negative discrimination value in such a way that lower frequency terms are produced from higher frequency components, with correspondingly better discrimination values [28], [29]. Specifically, if the high frequency nondiscriminators are taken in groups, and "phrases" are formed for cooccurring sets of nondiscriminators, the phrases will obviously exhibit lower document frequencies than the original components. The process is illustrated in the example of Fig. 12, for two original high frequency terms T_i and T_j, exhibiting an area of overlap consisting of the documents to which both terms are assigned. The frequency range of T_i and T_j may be reduced, by assigning term T_i' to those documents in which T_i only appears but not T_j; similarly T_j' is assigned to items in which only T_j was originally present, while the phrase T_{ij} is assigned to documents originally containing both terms.

The transformation illustrated in Fig. 12 may be generalized by using larger term groups (phrases with more than two components), obtained for example through an automatic term clustering process. These phrases can then be assigned

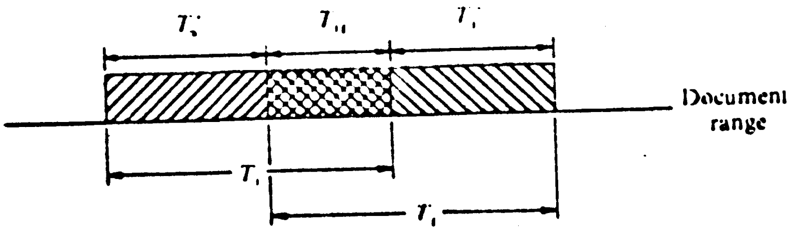


FIG. 12. *Illustration for generation of low frequency term combinations.*

to documents and queries whenever the corresponding components are present in addition to, or instead of, the original high-frequency terms. The expense of a term clustering process can be avoided entirely by simply taking the high-frequency terms occurring in sample user queries or documents, and defining term pairs, triples, quadruples, etc., for certain cooccurring terms.

One particular phrase formation process, tested experimentally, consists in arranging the nondiscriminators occurring in user queries in increasing discrimination order (worst nondiscriminator first), and arbitrarily defining for each set of three adjacent nondiscriminators three term pairs and one term triple [29]. The process is illustrated in Table 18, where it is seen that a single pair is formed from two original nondiscriminators; three pairs and a triple are formed from 3 terms, 5 pairs and 2 triples are produced from 4 terms; 6 pairs and 2 triples from 5 and 6 terms, and so on.[4]

TABLE 18
Experimental phrase formation procedure

High frequency nondiscriminators in queries	Newly defined phrases
2 terms $\underline{T_i T_j}$	T_{ij}
3 terms $\underline{T_i T_j T_k}$	$T_{ij}, T_{ik}, T_{jk}, T_{ijk}$
4 terms $\underline{T_i T_j T_k} T_l$	$T_{ij}, T_{ik}, T_{jk}, T_{ijk}$ T_{jl}, T_{kl}, T_{jkl}.
5 terms $\underline{T_i T_j T_k} T_l T_m$	$T_{ij}, T_{ik}, T_{jk}, T_{ijk}$ $T_{kl}, T_{km}, T_{lm}, T_{klm}$
6 terms $\underline{T_i T_j T_k}\ \underline{T_l T_m T_n}$	$T_{ij}, T_{ik}, T_{jk}, T_{ijk}$ $T_{lm}, T_{ln}, T_{mn}, T_{lmn}$

For the three sample collections used previously, an average number of 8.6, 2.16, and 10.8 new term pairs and triples are generated from the nondiscriminators for each document in the CRAN, MED, and Time collections, respectively, by the foregoing process. The document frequency distribution for the simple term nondiscriminators used in the phrase generation process is shown in Table 19 together with the distribution for the corresponding pairs and triples. It is obvious from Table 19 that as expected the average document frequency is much higher for singles than for pairs, and for pairs than for triples.

The newly generated phrases can be assigned to documents and queries in various combinations. Singles, pairs, and triples can all be used together (SPT);

[4] In a practical implementation, the phrase formation model of Table 18 need not of course be followed precisely. In fact, it is unnecessary physically to form any phrases at all; instead in each query or document, the high-frequency nondiscriminators can be flagged appropriately, and the formation of the corresponding pairs and triples can be made implicitly. When query and document vectors are compared in a retrieval situation, the matching coefficients between the vectors are simply adjusted to account for the presence of matching phrases.

TABLE 19
Document frequency distribution for high frequency nondiscriminators used in phrase generation

	Document frequency range	Single terms	Term pairs	Term triples
	0		0	1
	1-9	0	6	12
	10-19	0	20	6
	20-29	0	13	2
	30-39	0	8	2
	40-49	0	6	2
CRAN	50-59	15	11	1
424	60-69	5	5	0
	70-79	9	2	1
	80-89	4	6	0
	90-99	4	1	0
	100-129	17	3	0
	130-159	14	0	0
	over 160	13	0	0
	0	—	6	14
	1-9	0	69	16
	10-19	3	13	0
	20-29	17	2	0
	30-39	33	0	0
	40-49	11	0	0
MED	50-59	9	0	0
450	60-69	8	0	0
	70-79	0	0	0
	80-89	3	0	0
	90-99	4	0	0
	100-129	0	0	0
	130-159	2	0	0
	over 160	0	0	0
	0	—	0	0
	1-9	0	4	9
	10-19	0	18	10
	20-29	0	17	4
	30-39	0	16	6
	40-49	8	7	2
Time	50-59	15	7	0
425	60-69	3	8	1
	70-79	8	7	0
	80-89	13	3	0
	90-99	10	2	0
	100-129	7	3	0
	130-159	10	0	0
	over 160	22	0	0

alternatively, pairs and triples can be added to the vectors, and the corresponding singles deleted (PT): pairs only could be added while deleting the corresponding singles (P): and so on. It is found experimentally that when the high-frequency nondiscriminators are used for phrase generation purposes, the PT method offers a high standard of performance [29]. The phrase generation process can however also be implemented by using as starting single terms the medium-frequency discriminators. In that case, the SPT process which preserves the single term discriminators in the document and query vectors is best.

The effectiveness of the right-to-left phrase generation method is demonstrated by the recall-precision output of Tables 20 and 21. Table 20 shows average precision values at ten recall points for phrase runs SPT, PT, ST and P: a control run using standard term frequency weighting but no phrases is also included. Results are shown separately for phrases obtained from the high-frequency nondiscriminators and from the medium frequency discriminators. The best results in each section of Table 20 are emphasized by a vertical bar alongside the precision values.

It may be seen from Table 20, that when the high-frequency nondiscriminators are combined into phrases, improvements over the standard *TF* run are obtained almost everywhere. The best runs are the PT and P runs, where the single term nondiscriminators are deleted when the phrases are introduced into the vectors. Substantial improvements are also obtained for the phrases derived from the discriminators, listed on the right-hand side of Table 20. However, in that case, the good runs are the SPT and ST runs in which the single term discriminators are maintained.[5]

A combined run in which the phrases obtained from the nondiscriminators are applied using the PT strategy, whereas phrases from discriminators are used with the SPT system is shown in the middle of Table 21, designated as PT + SPT. This phrase procedure is compared against the previously mentioned optimum single term weighting process, labelled ($f_i^k \cdot IDF_k$) (term frequency multiplied by inverse document frequency). The best results are again emphasized by a vertical bar. It is seen that the single term weighting process is somewhat preferable for the CRAN collection: however, the phrase generation methods are superior both for MED and Time.[6]

The effectiveness of the vocabulary improvement obtained from the phrase generation procedure is summarized by the statistical significance output of Table 22. For each of the three collections the following pairs of runs are compared:

(a) term frequency f_i^k run against PT phrase run using nondiscriminators;
(b) f_i^k run against SPT phrase run using discriminators;
(c) f_i^k run against combined PT + SPT; and
(d) combined PT + SPT against combined $f_i^k \cdot IDF$ weighting.

The results of Table 22 show that only for two comparisons using the CRAN collection does the phrase process not perform as expected. In all other cases, the

[5] The elimination of the single term nondiscriminators is obviously useful, whereas the elimination of the single term discriminators would bring about considerable losses.

[6] The $f^k \cdot IDF_k$ weighting system can of course be applied in addition to the phrases.

TABLE 20

Average precision values at indicated recall points for three collections

Collection	Recall	Standard term frequency weights (TF)	Phrases formed from high frequency nondiscriminators: SPT	PT	ST	P	Phrases formed from medium frequency discriminators: SPT	PT	ST	P
CRAN 424	.1	.6844	.6293	.6620	.6787	.6564	.6917	.4737	.6595	.4582
	.2	.5303	.4797	.5283	.5324	.5404	.5536	.3145	.5087	.2970
	.3	.4689	.4242	.4337	.4694	.4820	.4977	.2740	.4748	.2711
	.4	.3482	.3336	.3430	.3455	.3620	.3787	.2224	.3508	.2106
	.5	.3134	.2903	.3000	.3092	.3106	.3532	.2067	.3134	.1825
	.6	.2556	.2366	.2426	.2529	.2460	.2931	.1697	.2625	.1475
	.7	.1989	.1879	.1942	.1978	.1994	.2176	.1175	.1998	.1152
	.8	.1631	.1572	.1595	.1598	.1590	.1802	.0973	.1617	.0952
	.9	.1265	.1270	.1345	.1272	.1360	.1430	.0813	.1303	.0796
	1.0	.1176	.1198	.1284	.1182	.1299	.1331	.0764	.1217	.0742
MED 450	.1	.7891	.7465	.8609	.8055	.8578	.8223	.6896	.8029	.6896
	.2	.6750	.6705	.7609	.6786	.7652	.7168	.5386	.6733	.5186
	.3	.5481	.5629	.6345	.5587	.6303	.5707	.4529	.5464	.4525
	.4	.4807	.4999	.5947	.4928	.5905	.5191	.3789	.4767	.3673
	.5	.4384	.4599	.5489	.4497	.5430	.4688	.3242	.4378	.3153
	.6	.3721	.3761	.4889	.3885	.4815	.3807	.2606	.3775	.2606
	.7	.3357	.3371	.4348	.3552	.4170	.3455	.2329	.3411	.2329
	.8	.2195	.2366	.3011	.2273	.3022	.2377	.1469	.2377	.1469
	.9	.1768	.1880	.2033	.1839	.2047	.1985	.1051	.1985	.1051
	1.0	.1230	.1229	.1427	.1213	.1440	.1229	.0914	.1219	.0914
Time 425	.1	.7496	.7744	.8471	.7545	.8274	.7654	.6307	.7589	.5987
	.2	.7071	.7366	.7952	.7151	.7766	.7654	.6251	.7159	.5712
	.3	.6710	.6708	.7539	.6760	.7586	.7144	.5546	.6853	.5353
	.4	.6452	.6357	.7254	.6431	.7255	.6909	.5017	.6509	.4617
	.5	.6351	.6347	.6732	.6326	.6907	.6644	.4662	.6408	.4377
	.6	.5866	.5859	.6320	.5888	.6363	.6105	.4438	.5922	.4162
	.7	.5413	.5354	.5897	.5482	.5945	.5726	.3987	.5567	.3663
	.8	.5004	.4924	.5320	.5137	.5462	.5355	.3539	.5161	.3263
	.9	.3865	.3996	.3997	.3934	.4038	.4289	.2147	.4069	.2050
	1.0	.3721	.3830	.3862	.3787	.3854	.4155	.1995	.3934	.1911

TF Standard term frequency weighting (word stem run)
SPT Single terms, pairs and triples used in queries and documents.
PT Pairs and triples used, corresponding single terms deleted
ST Single terms retained, triples added
P Pairs added, corresponding single terms deleted

phrase methods produce significant improvements over the standard f_i^k weighting for single terms, and they are also superior to the $f_i^k \cdot IDF$ combined term weighting system.

C. *Left-to-right thesaurus transformation.* The left-to-right transformation takes low frequency terms and transforms them into units of higher frequency by grouping a number of the low-frequency entities into classes. The term classes are then characterized by frequency properties equivalent to the sum of the frequencies of the individual components.

The classical way of combining individual terms into classes is by means of a *thesaurus*. Such a thesaurus specifies a grouping of the vocabulary, where items included in the same class are normally considered to be related in some sense—for example, by being synonymous, or by exhibiting closely similar content characteristics. Obviously, if a number of low frequency terms are grouped to form

TABLE 21

Average precision values at indicated recall points for phrase processing

Collection	Recall	Standard term frequency run (*TF*)	Best phrase process *PT* + *SPT*	Best frequency weighting (*TF* · *IDF*)
	.1	.6844	.7311	.7573
	.2	.5303	.6227	.6241
	.3	.4689	.5404	.5348
	.4	.3482	.4387	.4457
CRAN 424	.5	.3134	.3594	.3935
	.6	.2556	.3054	.3182
	.7	.1989	.2426	.2521
	.8	.1631	.1780	.1953
	.9	.1265	.1490	.1388
	1.0	.1176	.1316	.1277
	.1	.7891	.8876	.8459
	.2	.6750	.8223	.7557
	.3	.5481	.6814	.6584
	.4	.4807	.6379	.5442
MED 450	.5	.4384	.5951	.4873
	.6	.3721	.5246	.4254
	.7	.3357	.4755	.3833
	.8	.2195	.3364	.2622
	.9	.1768	.2420	.2123
	1.0	.1230	.1742	.1469
	.1	.7496	.8860	.8536
	.2	.7071	.7964	.7901
	.3	.6710	.7761	.7568
	.4	.6452	.7461	.7305
Time 425	.5	.6351	.7020	.6783
	.6	.5866	.6563	.6243
	.7	.5413	.6010	.5823
	.8	.5004	.5483	.5643
	.9	.3865	.4231	.4426
	1.0	.3721	.4118	.4170

TF Standard term frequency weighting (word stem run)

PT + *SPT* Use pairs and triples derived from nondiscriminators plus singles, pairs and triples obtained from discriminators.

TF · *IDF* Use a term weight consisting of term frequency multiplied by the inverse document frequency

TABLE 22

Statistical significance output for selected runs of Table 21 (probability that run B is significantly better than run A, except where A > B indicates that test is made in reverse direction)

	CRAN 424		MED 450		Time 425	
	t-test	Wilcoxon	t-test	Wilcoxon	t-test	Wilcoxon
A. Standard f_i^k run vs. B. PT phrases from nondiscriminators	0.18 (A > B)	0.41	0.00	0.00	0.00	0.00
A. Standard f_i^k run vs. B. SPT phrases from discriminators	0.00	0.00	0.00	0.00	0.00	0.00
A. Standard f_i^k run vs. B. Combined PT + SPT phrases	0.02	0.00	0.00	0.00	0.00	0.00
A. $f_i^k \cdot IDF$ weights vs. B. Combined PT + SPT phrases	0.01 (A > B)	0.00	0.00	0.00	0.78	0.81

a thesaurus class, the class will exhibit a much higher document frequency, and most likely a better discrimination value, than any of the original terms.

There exist well-known procedures for constructing thesauruses either manually or automatically [10], [12], [24]. In the latter case, automatic term classification methods may be used to generate the appropriate term groups [30]. According to the theory presented earlier, the main virtue of a thesaurus is the classification of low frequency terms into higher frequency classes. The corresponding class identifiers can then be incorporated into query and document vectors in addition to, or instead of, the individual term components.

To test this theory, it is in principle necessary to construct new thesauruses for the three test collections used experimentally, and to impose appropriate frequency restrictions on the input vocabulary. A shortcut method can be used for experimental purposes which consists in using available term classifications for each of the three subject areas under consideration (aerodynamics, medicine, and world affairs), while deleting from the existing term classes entries whose document frequency exceeds a given threshold. The resulting thesaurus classes are not directly comparable to classes obtained by using only the low frequency terms for clustering purposes. However, the experimental recall-precision results may be close to those produced by the alternative, possibly preferred, methodology.

The document frequency cutoff actually used for deciding on inclusion of a given term in the experimental thesauruses was 19, 15, and 19 for the CRAN, MED, and Time collections respectively; that is, terms with document frequencies smaller than or equal to the stated frequencies were included. For the three test collections, the process creates 19, 60, and 26 thesaurus classes, respectively. The document frequency distributions of the rare terms included in the thesauruses and of the corresponding thesaurus classes are shown in Table 23.

A comparison of the document frequency ranges in the two main columns of Table 23 makes it clear that the thesaurus classes in the right-most column exhibit much higher frequency characteristics than the original terms. Furthermore, when the document frequency ranges of the thesaurus classes are compared with the frequency ranges of the good discriminators in the middle column of Table 17 (that is, 20–40 for CRAN, 5–20 for MED, and 5–30 for Time), it appears that the majority of the thesaurus classes fall into the desired frequency range.

The recall-precision results obtained with the low-frequency term classification is shown in column 3 of Table 24, labelled "thesaurus". In each case, a thesaurus class identifier was added to a document or query vector with a basic weight of 1, whenever one of the terms included in that thesaurus class was originally present in the document or query. A comparison between columns 2 and 3 of Table 24, reflecting the performance of the basic word stem indexing method with term frequency weighting (f_i^k), and the thesaurus process consisting of word stem plus thesaurus classes makes it obvious that the thesaurus process is much superior. Moreover, the differences in performance are statistically significant as shown in the last row of Table 25.

The performance of a combined left-to-right (thesaurus) and right-to-left (phrase) transformation process is shown in columns 4 and 5 of Table 24. Column 4 contains the output for "thesaurus plus PT phrases", where pairs and triples are derived from high-frequency nondiscriminators only. The next column, labelled "thesaurus plus PT + SPT", uses phrases derived both from discriminators as well as from nondiscriminators. For comparison purposes, the output corresponding to the best phrase process and best frequency weight method from Table 21 is copied again in Table 24.

The performance of the best indexing method of any of those reviewed in the current study is emphasized by a double bar in Table 24. It is seen that the results in the last three columns of the table covering best frequency weighting, best phrase, and best combined phrase and thesaurus method do not differ widely, except for the MED collection where statistically significant advantages are apparent for thesaurus and phrases. However, for all three collections, the combined thesaurus plus phrase process gives the best overall performance; and that performance is normally at least twenty percent better than the single term (word stem) term frequency (f_i^k) or binary weight (b_i^k) control run. A graphic illustration of the performance differences for the three experimental collections is shown in the recall-precision plots of Fig. 13.

At the present time, no automatic indexing methodology is known which would improve upon the performance of the combined thesaurus plus phrase methods generated from the indexing theories included in this study.

TABLE 23
Document frequency distribution of rare terms used for thesaurus construction

	Document frequency range	Rare terms used for thesaurus	Document frequency range	Thesaurus classes created by process
CRAN	1-3	3	1-5	3
	4-6	6		
	7-9	4	6-10	3
	10-12	3		
	13-15	2	11-15	4
	16-19	4	16-20	2
	20+	0	21-25	4
			26-30	0
			31-35	3
			36-40	0
MED	1-3	14	1-5	14
	4-6	15		
	7-9	8	6-10	16
	10-12	17		
	13-15	12	11-15	21
	16-19	0	16-20	5
	20+	0	21-25	4
			26-30	0
			31-35	0
			36-40	0
Time	1-3	2	1-5	1
	4-6	3		
	7-9	4	6-10	6
	10-12	7		
	13-15	8	11-15	5
	16-19	5	16-20	8
	20+	0	21-25	3
			26-30	2
			31-35	0
			36-40	1

TABLE 24

Recall precision output for thesaurus processing

	R	Standard term freq f_i^k	Thesaurus	Thesaurus + PI phrases (mendised)	Thesaurus + PI + SPT phrases	Best phrase process PI + SPT	Best freq weight $f_i^k \cdot IDF_k$
CRAN	0.1	.6844	.7463	7129	7614	7311	.7573
	0.2	.5303	.5806	5720	6887	.6227	6241
	0.3	.4689	.5052	4793	5574	5405	.5348
	0.4	.3482	3811	3738	4664	4387	4457
	0.5	.3134	3375	3240	3954	3594	3935
	0.6	.2556	.2755	.2732	3252	.3054	.3182
	0.7	.1989	.2316	2279	2572	2426	2521
	0.8	.1631	.1885	.1842	1803	1780	1953
	0.9	.1265	.1375	.1433	1486	1490	1388
	1.0	.1176	.1282	.1387	1327	1316	1277
MED	0.1	.7891	.8319	.8712	8867	8876	8459
	0.2	.6750	.7283	.7766	8199	8223	7557
	0.3	.5481	.6151	.6556	6948	6814	6584
	0.4	.4807	.5371	.6121	6334	.6379	.5442
	0.5	.4384	.4741	.5660	6067	.5951	4873
	0.6	.3721	.4193	4896	.5318	.5246	4254
	0.7	3357	.3832	4594	.5035	.4755	.3833
	0.8	.2195	.2819	3463	.3844	.3364	2622
	0.9	.1768	2267	2694	.3070	.2420	.2123
	1.0	.1230	1640	1791	2074	1742	.1469
Time	0.1	.7496	.7392	8649	8761	.8860	.8536
	0.2	.7071	.7166	7984	7972	7984	.7901
	0.3	.6710	6935	7631	.7778	.7761	.7568
	0.4	.6452	.6627	7258	.7465	.7461	.7305
	0.5	.6351	.6541	6821	7027	.7020	.6783
	0.6	.5866	6070	6358	6524	.6563	.6243
	0.7	.5413	.5598	5930	.6010	6010	.5823
	0.8	.5004	5111	.5421	.5523	.5483	.5643
	0.9	.3865	4091	.4185	.4260	.4231	4426
	1.0	3721	.3950	.4040	.4149	.4118	.4170

A number of questions remain for further examination. The following are the most important for a practical application of the theory:

(a) To what extent can one justify the replacement of the complicated discrimination value computations by the simple document frequency model?

(b) Can the computation of term values obtained from a static model of a given document collection be maintained in a dynamic environment where old documents are removed, and new ones are added? If not, how often must one recompute the term values?

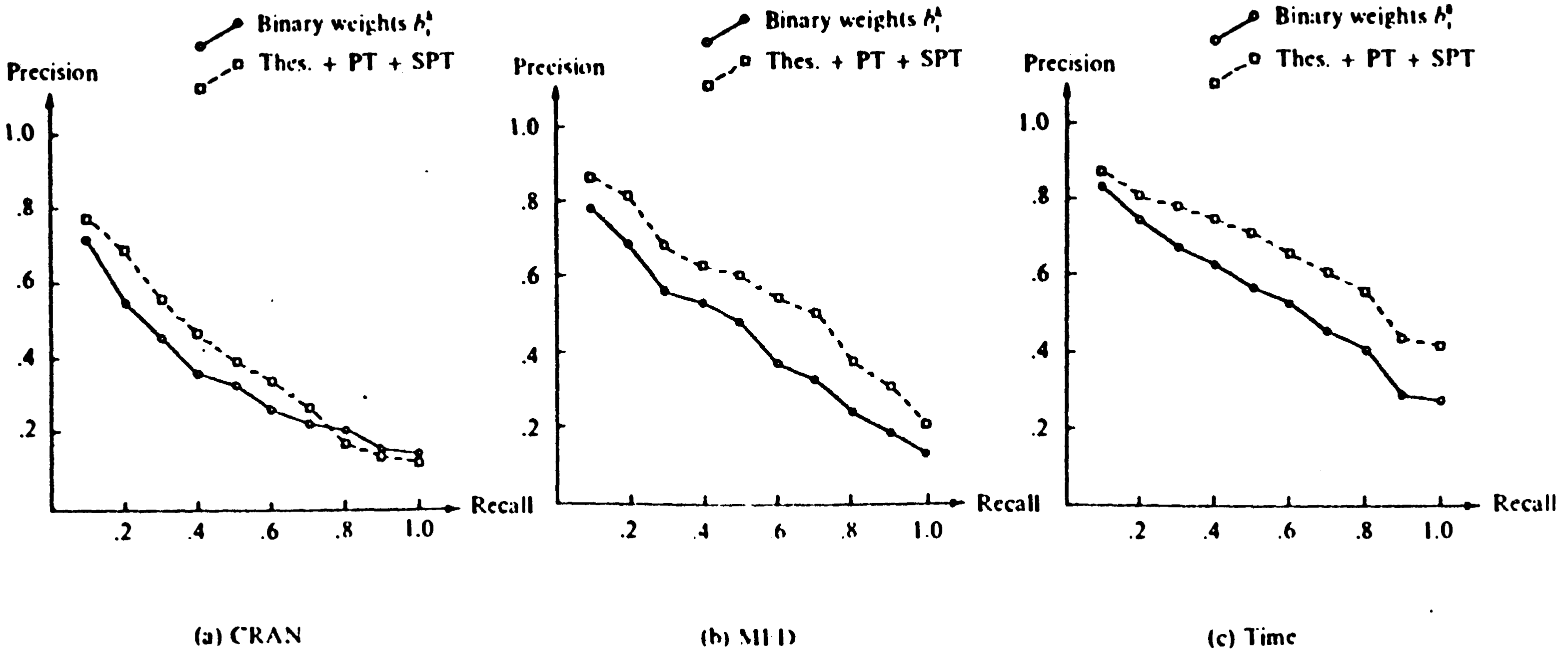


FIG. 13. *Comparison of standard word stem indexing with binary weights and combined left-to-right and right-to-left transformation (thesaurus plus phrases)*

TABLE 25

Statistical significance output for runs of Table 24 (all tests for run A > B)

	CRAN		MED		Time	
	t-test	Wilcoxon	t-test	Wilcoxon	t-test	Wilcoxon
A. Thesaurus + PT + SPT phrases B. $f^k \cdot IDF_k$ weights	.8085	.9855	.0000	.0000	.6874	.6833
A. Thesaurus + PT + SPT phrases B. PT + SPT phrases	.0000	.0013	.0000	.0022	.4524	.9657
A. Thesaurus B. Standard term frequency f_i^k	.0000	.0000	.0000	.0000	.0000	.0003

(c) Can the term values obtained from a collection in a given subject area be used for collections in different subject areas?

Questions relating to dynamic collection and thesaurus maintenance have been examined elsewhere [31], [32]. They must be related to the current indexing theory if a practical implementation is contemplated.

REFERENCES

[1] K. SPARCK JONES, *A statistical interpretation of term specificity and its application in retrieval*, J. Documentation, 28 (1972), pp. 11-21.

[2] P. ZUNDE AND V. SLAMECKA, *Distribution of indexing terms for maximal efficiency of information transmission*, Amer. Documentation, 18 (1967), pp. 106-108.

[3] H. P. LUHN, *A statistical approach to mechanized encoding and searching of literary information*, IBM J. Res. Develop., 1 (1957), pp. 309-317.

[4] ——, *The automatic derivation of information retrieval encodements for machine readable texts*, Information Retrieval and Machine Translation, Part 2, A. Kent, ed., Interscience, New York, 1961.

[5] C. E. SHANNON, *A mathematical theory of communication*, Bell Systems Tech. J., 27 (1948), pp. 379-423, 623-656.

[6] F. J. DAMERAU, *An experiment in automatic indexing*, Amer. Documentation, 16 (1965), pp. 283-289.

[7] S. F. DENNIS, *Law, language, words, entropy, and automatic indexing*, unpublished manuscript.

[8] ——, *The design and testing of a fully automatic indexing-searching system for documents consisting of expository text*, Information Retrieval: A Critical Review, G. Schecter, ed., Thompson Book Co., Washington, 1967, pp. 67-94.

[9] K. BONWIT AND J. ASTE TONSMAN, *Negative Dictionaries*, Scientific Rep. ISR-21, Section VI, Department of Computer Science, Cornell University, Ithaca, N.Y., October 1970.

[10] G. SALTON, *Experiments in automatic thesaurus construction for information retrieval*, Proc. IFIP Congress 71, Ljubljana, North Holland Publishing Co., Amsterdam, 1972.

[11] C. R. SAGE, R. R. ANDERSON AND P. F. FITZWATER, *Adaptive information dissemination*, Amer. Documentation, 16 (1965), pp. 185–200.

[12] G. SALTON, *Automatic Information Organization and Retrieval*, McGraw-Hill, New York, 1968.

[13] ———, *A new comparison between conventional indexing (Medlars) and automatic text processing (SMART)*, J. ASIS, 23 (1972), No. 2, pp. 75–84.

[14] V. E. GIULIANO AND P. E. JONES, *Linear associative information retrieval*, Vistas in Information Handling, P. Howerton, ed., Spartan Books, Washington, D.C., 1963.

[15] L. B. DOYLE, *Indexing and abstracting by association*, Amer. Documentation, 13 (1962), pp. 378–390.

[16] H. E. STILES, *The association factor in information retrieval*, J. ACM, 8 (1961), pp. 271–279.

[17] M. E. MARON AND J. L. KUHNS, *On relevance, probabilistic indexing and information retrieval*, Ibid., 7 (1960), pp. 216–244.

[18] M. E. MARON, *Automatic indexing: an experimental inquiry*, Ibid., 8 (1961), pp. 404–417.

[19] N. HOUSTON AND E. WALL, *The distribution of term usage in manipulative indexes*, Amer. Documentation, 15 (1964), pp. 105–114.

[20] E. WALL, *Further implications of the distribution of index term usage*, Proc. Annual Meeting of the American Documentation Institute, 1 (1964), pp. 457–466.

[21] J. C. COSTELLO AND E. WALL, *Recent improvements in techniques for storing and retrieving information*, Studies in Coordinate Indexing, 5, Documentation Inc., Washington, D.C., 1959.

[22] H. L. RESNIKOFF AND J. L. DOLBY, *Access: A study of information storage and retrieval with emphasis on library information systems*, Interim Report, R. and D. Consultants, Los Altos, California, May 1971.

[23] H. L. RESNIKOFF, *On information systems with emphasis on the mathematical sciences*, Conference Board of Mathematical Sciences, Washington, January, 1971.

[24] G. SALTON AND M. E. LESK, *Computer evaluation of indexing and text processing*, J. ACM, 15 (1968), pp. 8–36.

[25] R. W. CRAWFORD, *Negative Dictionary Construction*, Scientific Rep. ISR-22, Section IV, Department of Computer Science, Cornell University, Ithaca, N.Y., November 1974.

[26] G. SALTON AND C. S. YANG, *On the specification of term values in automatic indexing*, J. Documentation, 29 (1973), pp. 351–372.

[27] A. WONG, R. PECK AND A. VAN DER MEULEN, *An adaptive dictionary in a feedback environment*, Scientific Rep. ISR-21, Section XIV, Department of Computer Science, Cornell University, Ithaca, N.Y., 1972.

[28] G. SALTON AND C. T. YU, *On the construction of effective vocabularies for information retrieval*, SIGPLAN SIGIR Symposium on Programming Languages and Information Retrieval, Gaithersburg, Maryland, November 1973.

[29] G. SALTON, C. S. YANG AND C. T. YU, *Contributions to the theory of indexing*, Information Processing 74, North Holland Publishing Co., Amsterdam, 1974, pp. 584–590.

[30] K. SPARCK JONES, *Automatic Keyword Classifications*, Butterworths, London, 1971.

[31] G. SALTON, *Dynamic document processing*, ACM Comm., 15 (1972), pp. 658–668.

[32] ———, *Proposals for a dynamic library*, *Information—Part 2*, 2 (1973), No. 3, pp. 5–27.