EFFECTIVE INFORMATION RETRIEVAL USING TERM ACCURACY

C.T. Yu* and G. Salton+

TR 75-249

July 1975

Department of Computer Science
Cornell University
Ithaca, New York 14853

*Department of Computing Science, University of Alberta, Edmonton, Alberta.

+Department of Computer Science, Cornell University, Ithaca, New York 14853.

This study was supported in part by the Canadian Research Council and in part by the National Science Foundation under grant GJ 43505.

Effective Information Retrieval Using Term Accuracy

C.T. Yu* and G. Salton+

Abstract

The performance of information retrieval systems can be evaluated in a number of different ways. Much of the published evaluation work is based on measuring the retrieval performance of an average user query. Unfortunately, formal proofs are difficult to construct for the average case.

In the present study, retrieval evaluation is based on optimizing the performance of a specific user query. The concept of query term accuracy is introduced as the probability of occurrence of a query term in the documents relevant to that query. By relating term accuracy to the frequency of occurrence of the term in the documents of a collection it is possible to give formal proofs of the effectiveness with respect to a given user query of a number of automatic indexing systems that have been used successfully in experimental situations. Among these are inverse document frequency weighting, thesaurus construction, and phrase generation.

1. The Term Discrimination Model

A good deal is known about the evaluation of automatic indexing and content analysis systems. Typically, a given process — for example, a term weighting

* Department of Computing Science, University of Alberta, Edmonton, Alberta.

+ Department of Computer Science, Cornell University, Ithaca, New York 14853.

This study was supported in part by the Canadian Research Council and in part by the National Science Foundation under grant GJ 43505.

system, or a phrase formation procedure — is applied to the documents of a collection and to a set of available user queries. The queries are then processed against the documents of the collection with and without the use of the indexing component under consideration. In either case, recall and precision figures are averaged over the number of available user queries, and the better system is judged to be the one with the higher average precision for stated levels of the recall.[+]

When recall and precision are averaged over a number of user queries, it is difficult to generate formal proofs of the effectiveness of system components because the relative frequency of occurrence of the many possible user queries is difficult to estimate. Experimental evidence can, however, be obtained of the retrieval effectiveness of various processes by using document collections in several subject areas, together with various types of user queries. When the experimental results show similar trends in a number of unrelated environment the presumption is strong that the experimental results are generally valid. [1]

One automatic indexing theory that has been thoroughly tested and found valuable in practice is the term discrimination model. [2,3] In this model, a good term is assumed to be one which when assigned as an index term to a collection of documents will render the documents as dissimilar as possible;

[+] Recall is the proportion of relevant items retrieved, and precision is the proportion of retrieved items that are relevant. Ideally one would like to obtain high recall by retrieving most everything that is relevant, as well as high precision by rejecting everything that is extraneous.

that is, it will cause the greatest possible separation between the documents in the indexing space.[+] Contrariwise, a poor term is one which renders the documents more similar, and therefore makes it harder to distinguish one document from another. The idea is that the greater the separation between individual documents, and the more dissimilar the respective index term sets (vectors), the easier it will be to retrieve some items while rejecting others; when, on the other hand, the documents exhibit similar term vectors — that is, when the document indexing space is bunched up — it will be impossible to insure the proper discrimination between relevant and nonrelevant items.

For each potential index term assigned as a content identifier to a collection of documents, a discrimination value (DV) can then be computed as a function of the difference between the space "densities" before and after the assignment of that term. The greater the difference in space densities, the more the space will have spread (because the documents will have become more dissimilar), and therefore the better the particular term will function as a discriminator. The space density may be defined as the average pairwise similarity between the documents of the collection, or more efficiently as the sum of the similarity coefficients between each document and a dummy "centroid" document defining the center of the space. [2,3]

[+] It is assumed that the similarity between two documents (or between a query and a document) can be ascertained by computing a coefficient of nearness between them based on the similarities between the respective term sets or "vectors".

The best discriminators generally exhibit positive discrimination values (the density of the document space decreases when the term is assigned); correspondingly, the worst discriminators have negative discrimination values because the documents become more similar to each other, and space density increases. In the middle, a large number of documents may be found whose assignment does not affect the space density; the corresponding discrimination values remain near zero. When the terms are arranged in decreasing order of their discrimination values, the ordering produces a rank for each term, known as the discrimination rank (DR). The largest discrimination value then corresponds to the lowest rank.

A study of a number of document collections in different subject areas reveals a relationship between the discrimination value of an index term and its <u>document frequency</u> (DF), defined as the number of documents in the collection to which the term is assigned. The following relationships appear to be generally valid [3]:

a) terms with very low document frequencies that may be assigned to one or two documents only are generally poor discriminators with discrimination ranks in excess of t/2 for a total of t existing terms;

b) terms with high DF, comprising those that are assigned to more than ten percent of the documents in a collection are the worst discriminators, with average discrimination ranks near t;

c) the best discriminators are those whose DF is neither too high nor too low, with document frequencies between n/100 and n/10 for n documents; their average discrimination ranks are generally below t/5 for t terms.

The first of these relationships may perhaps seem counterintuitive: one would expect that terms occurring with extreme rarity might be the best discriminators — obviously they should be usable to distinguish the few documents in which they occur from the remainder. The problem is that their presence, or absence, affects only very few items out of many thousands; thus the density of the whole document space measured by the average similarity between all items is hardly affected by the rare terms. The retrieval performance averaged over many user queries is similarly unaffected by the presence, or absence, of the rare terms since these terms are included at best in a very small number of user queries.

The discrimination value theory gives rise to an automatic indexing strategy in which good discriminators are used directly for content identification purposes, whereas bad discriminators — those whose document frequencies are either too high, or too low — are transformed into better discriminators by changing their frequency characteristics [3]:

a) Terms with average document frequencies (between approximately n/100 and n/10 for n documents) are used directly for indexing purposes; these terms include the vast majority of the good discriminators.

b) Terms whose document frequency is too high — above n/10 — comprise the worst discriminators. These terms are too general in nature, or too broad, to permit proper discrimination among the documents; hence their use produces an unacceptable precision loss (it leads to the retrieval of too many extraneous items). These terms are transformed into lower frequency entities by using them in pairs (or triples) as indexing phrases. The frequency of assignment of a phrase is generally smaller than the assignment frequencies of the respective phrase componenets, and the phrase discrimination value is correspondingly greater. In systems using weighted terms in which the weight functions as an index of term importance, the phrase weight might be defined as the *average* of the weights of the phrase components.

c) Terms whose document frequency is too low — below n/100 — are so rare and specific that they cannot retrieve an acceptable proportion of the documents relevant to a query; hence their use depresses the recall performance. These terms are transformed into higher frequency entities by including them in *term classes*, and assigning these term classes as content identifiers instead of the individual terms.[+] A term class will generally exhibit a higher assignment frequency than a single element of a class with a corresponding greater discrimination value. The class weight may be taken to be the *sum* of the weights of the individual class elements.

The indexing strategy based on the discrimination value theory thus introduces indexing phrases generated from high-frequency components as a means for enhancing the precision performance; low-frequency components, on the other hand, are combined into term classes for recall improving purposes.

The use of the document frequency for the representation of term effectivne is a simplifying device which is generally, but not absolutely valid. In cases where weighted document and query vectors are used, that is, where each term is assigned a weight (such as for example its frequency of occurrence in a given document), a more accurate indication of term value may be obtained by looking at the term frequency (or term weight) distribution across the documents of a collection.

[+] Term classes are specified by thesauruses, or synonym dictionaries in many retrieval environments.

Consider in particular the set of terms exhibiting a common total occurrence frequency (or a common total weight) in a given collection. If only a single frequency indicator (such as the total frequency, or the total weight) were to be utilized as an indication of term value, all these terms might be assigned the same value for indexing purposes. It turns out, however, that terms with a flat distribution — those that have the same occurrence frequency, or the same weight, in every document in which they occur — are not nearly so effective for purposes of document discrimination as the terms whose frequency distribution is skewed. The latter may occur with high occurrence frequencies in some documents, and much more rarely in others.

This brings up a new term value parameter usable for systems based on weighted index terms. The best discriminators among the term set exhibiting a common total occurrence frequency, or a common total weight, are normally those whose document frequency is minimal. In other words, the best terms for content representation appear to be those whose occurrence frequency, or weight may be large in individual documents, but whose total assignment frequency over the whole document collection is small. [4] The frequency, or weight distribution for such terms will be skewed, because their presence is concentrated in only a few documents.

For weighted term assignments, an appropriate indication of term value is then the inverse document frequency (IDF) which varies inversely with the document frequency of the given term. Term value parameters based on a weighting function which includes the inverse document frequency have been extensively investigated and found to be easy to apply, and effective in retrieval. [4,5]

To summarize, the discrimination value theory assigns the highest value to terms with average document frequency and skewed weight (or frequency) distributions. Terms with excessively high document frequencies are replaced by phrases for precision enhancement; terms with very low document frequencies are on the other hand grouped into term classes to improve the recall. A good deal of experimental evidence exists to validate this theory, based in each case on averaging the performance over many user queries.

2. The Term Accuracy Model

The main conceptual difference between the retrieval model described earlier, and the one to be examined in the remainder of this study is that in the new system the evaluation of retrieval effectiveness will be based on performance optimization with respect to a particular query Q (rather than on the average performance over many user queries).

Consider, in particular, a user query Q. The performance criterion to be used is the ratio between the average similarity of the documents relevant to that query and that of the documents not relevant to the query. More specifically, let R and I denote the set of relevant documents and the set of nonrelevant documents respectively, corresponding to query Q. The desired performance criterion is then

$$\frac{E\{f(Q, D_1) \mid D_1 \epsilon R\}}{E\{f(Q, D_2 \mid D_2 \epsilon I\}} \qquad (1)$$

where E represents the expected value of the matching (similarity) function between Q and the document sets D_1 and D_2 respectively. It is clear that a large ratio for (1) implies that the relevant items are more similar to the query than the nonrelevant ones; hence the retrieval performance should be satisfactory. The reverse is true when the ratio of expression (1) is small.

For present purposes, the similarity function f is chosen as the simple vector matching function, that is, $f(X, Y) = \sum_{i=1}^{n} x_i y_i$, where $X = (x_1, x_2, \ldots, x_n)$ and $Y = (y_1, y_2, \ldots, y_n)$ represent document and query vectors. Furthermore, initially all document and query vectors are assumed to be binary (consisting of 0 and 1 vector elements only). For the original, unmodified document and query vectors, the similarity function f thus measures the number of matching vector elements (terms).

The indexing methods to be studied in this report are used in each case to modify the original query and document vectors Q and D, thereby replacing them by new, altered vectors Q* and D* respectively. If the particular indexing method under investigation is effective, then one may expect that the ratio (1) improves when applied to the modified situation, compared with the original. In other words, the presumption is that a useful indexing system brings the relevant items relatively closer to the query than the nonrelevant ones. Based on such a rationale, the following definition is proposed:

Definition: An indexing system is effective if

$$\frac{E\{f(Q^*, D_1^*) \mid D_1^* \varepsilon R\}}{E\{f(Q^*, D_2^*) \mid D_2^* \varepsilon I\}} \geq \frac{E\{f(Q, D_1) \mid D_1 \varepsilon R\}}{E\{f(Q, D_2) \mid D_2 \varepsilon I\}} \qquad (2)$$

where Q* and D* represent the modified query and document vectors, and Q and D the original ones.

Before actually considering the various indexing devices, it is necessary to characterize the usefulness of the query terms based on their frequency characteristics. Consider a query $Q = \{q_1, q_2, \ldots, q_m\}$ where q_i is the ith term assigned to query Q. The basic assumption to be made is that for any query term q_i, $1 \leq i \leq m$, an inverse relation exists between its document frequency and the probability that any document in which q_i occurs is relevant to query Q. In other words, one assumes that query terms which occur in many documents of the collection are not useful to distinguish the relevant items from the nonrelevant; rare terms on the other hand, that are assigned to very few documents have a good chance of occurring in the relevant ones.

More formally, consider query term q_i, $1 \leq i \leq m$, and let r_i and σ_i denote the number of documents containing q_i that are respectively relevant and nonrelevant to query Q. Using these parameters the accuracy of a query term q_i with respect to Q may be defined as its probability of occurrence in a document relevant to Q, that is, as $r_i/(r_i + \sigma_i)$.

The inverse relationship between term accuracy and document frequency may now be formalized as follows:

Assumption 1: Let q_i and q_j, $1 \leq i, j \leq m$, denote two distinct terms of query Q. Let r_i and r_j be the number of documents relevant to Q that contain q_i and q_j respectively; similarly, let σ_i and σ_j denote the number of documents nonrelevant to Q that contain q_i and q_j. If the document frequency of q_i is at least as high as that of q_j (that is, $r_i + \sigma_i \geq r_j + \sigma_j$), then one assumes that the accuracy of q_j is at least as high as that of q_i, or

$$\frac{r_i}{(r_i + \sigma_i)} \leq \frac{r_j}{(r_j + \sigma_j)} . \qquad (3)$$

It should be noted that the accuracy of a query term is not a property intrinsic to the term, as are document frequency, or query term weight, for example. Rather it is a term property with respect to some query Q. A given term may have as many values of accuracy as there are queries in which it appears, and its accuracy might be high with respect to some queries and low with respect to some others.+ This explains the apparent discrepancy between the experimental evidence cited earlier for the discrimination value model (that the best discriminators are terms of medium document frequency) and the current assumption that the best terms are the low frequency ones. Averaged over a set of user queries, the most useful terms exhibit everage document frequency; but <u>with respect to a specific query</u>, the query terms that are best capable of distinguishing the relevant items from the nonrelevant are those of low document frequency.

With these preliminaries, it is now possible formally to prove the effectiveness of the simple indexing devices that were previously shown effective by the discrimination value model, including inverse document frequency weighting, term class (thesaurus) generation for low-frequency terms, and indexing phrase transformation for high-frequency terms.

+ On the other hand, when two or more terms cooccur in several queries, the relative order of their accuracies is the same; that is, if the document frequency of term t_i is not smaller than that of t_j, then the accuracy of t_i with respect to any query Q containing both t_i and t_j is not larger than the corresponding accuracy for t_j.

3. Inverse Document Frequency Weighting

The inverse document frequency (IDF) weighting method leaves the actual terms assigned to a query or document unchanged, but adds term weights that are inversely related to the document frequencies of the terms. Thus, if query term q_i exhibits a higher document frequency then q_j, then $w_i \leq w_j$, where w_k is the weight of term k in the modified query Q*. Without loss of generality one may assume that in the modified query, the terms are listed in decreasing weight order, that is, $w_1 \geq w_2 \geq \dots \geq w_m$.

Based on these assumptions, the IDF weighting method is now shown to be an effective indexing system.

<u>Proposition 1</u>: Under the conditions of Assumption 1, the inverse document frequency weighting method is an effective indexing system.

<u>Proof</u>: From the definition of r_i, query term q_i is known to occur in r_i documents relevant to Q. Thus, q_i contributes a total of r_i to the matching function between Q and the relevant document set R, that is, $f(Q, \sum_{D_1 \varepsilon R} D_1)$. Similarly q_j contributes r_j, $1 \leq j \leq m$, to the same expression. From the bilinear property of f, one has

$$f(Q, \sum_{D_1 \varepsilon R} D_1) = \sum_{D_1 \varepsilon R} f(Q, D_1) = \sum_{i=1}^{m} r_i .$$

In the same way, one obtains $\sum_{D_2 \varepsilon I} f(Q, D_2) = \sum_{i=1}^{m} \sigma_i$. One concludes that

$$\frac{E\{f(Q, D_1) \mid D_1 \varepsilon R\}}{E\{f(Q, D_2) \mid D_2 \varepsilon I\}} = \frac{\sum_{i=1}^{m} r_i \;/\; |R|}{\sum_{i=1}^{m} \sigma_i \;/\; |I|} \tag{4}$$

where $|R|$ and $|I|$ represent the number of documents that are respectively relevant and nonrelevant to query Q.

When the document and query vectors are modified by assigning the IDF weights, a match between query term q_i and a relevant document produces an increase in the query-document similarity function of w_i^2. Proceeding as above, one now obtains

$$\frac{E\{f(Q^*, D_1^*) \mid D_1^* \varepsilon R\}}{E\{f(Q^*, D_2^*) \mid D_2^* \varepsilon I\}} = \frac{\sum_{i=1}^{m} w_i^2 r_i \;/\; |R|}{\sum_{i=1}^{m} w_i^2 \sigma_i \;/\; |I|} \tag{5}$$

By comparing (4) and (5), it is sufficient to show that

$$\left(\sum_{i=1}^{m} w_i^2 r_i\right) \cdot \left(\sum_{i=1}^{m} \sigma_i\right) \geq \left(\sum_{i=1}^{m} w_i^2 \sigma_i\right) \cdot \left(\sum_{i=1}^{m} r_i\right). \tag{6}$$

An expansion of the terms in (6) shows that (6) is true if and only if

$$\sum_{j=1}^{m-1} \left[\sum_{k=j+1}^{m} (w_j^2 - w_k^2)(r_j \sigma_k - r_k \sigma_j) \right] \geq 0. \tag{7}$$

But from the manner of assigning the IDF weights, one has $w_j^2 \geq w_k^2$, since $j \leq k$. Furthermore, by Assumption 1, $r_j/(r_j + \sigma_j) \geq r_k/(r_k + \sigma_k)$. Thus $r_j \sigma_k \geq r_k \sigma_j$.

Since every term in (7) is nonnegative, expression (7) is certainly satisfied, and the desired result (that (5) is not smaller than (4)) follows. The IDF weighting method is thus an effective indexing system.

4. Thesaurus (Term Class) Transformation

A standard information retrieval thesaurus may be used to group sets of semantically related terms into thesaurus classes. The thesaurus class weights are normally taken to be some function of the sum of the weights of the individua terms in the respective classes. For present purposes, a thesaurus method is introduced in which the number of terms per class is restricted to exactly two.

It was seen earlier that a thesaurus functions as a recall improving device. In particular, if a sufficient number of relevant documents cannot be retrieved in response to some query Q, the formation of query term classes may be expected to lead to additional relevant items, because new "related" terms are effectively introduced into document and query vectors when term classes are assigned instead of single terms. If, however, the terms included in a thesaurus class have low accuracy, then the addition of these terms to the queries is likely to lead to unacceptable losses in precision. For this reason, thesaurus classes are formed only for high accuracy terms.

More formally, let the medium accuracy for the terms in a specific query Q be defined as

$$\left(\sum_{i=1}^{m} r_i\right) / \left(\sum_{i=1}^{m} (r_i + \sigma_i)\right).$$

The query terms in Q defined as high accuracy terms are those whose accuracy is at least as large as the medium accuracy. (Note that the restricted applicabi used here for thesaurus construction is completely compatible with the evidence supplied by the discrimination value model. Since high accuracy implies low document frequency by Assumption 1, the high-accuracy thesaurus to be constructed here is effectively equivalent to the low-frequency term class generation indicated by the earlier DV model).

The thesaurus construction method to be examined provides for the grouping of "semantically related" high-accuracy terms. Two terms are assumed to be semantically related if they refer to similar objects in the same context. In a retrieval environment, the context is represented by a user query, and the objects are the documents. To be semantically similar, two terms must relate the documents to a given query in similar ways. Thus, "semantic relationship" like accuracy are term properties that may change from query to query.

It is now possible to show that the modified thesaurus construction method applied to high-accuracy terms yields an effective retrieval system.

Proposition 2: Let q_j be a term in query Q where

$$r_j/(r_j + \sigma_j) \geq \left(\sum_{i=1}^{m} r_i\right) / \left(\sum_{i=1}^{m} (r_i + \sigma_i)\right),$$

and let q_k be a term not present in Q, but semantically related to q_j. If the accuracy of q_k is not smaller than that if q_j, that is, if $r_k/(r_k + \sigma_k) \geq r_j/(r_j + \sigma_j)$, then the combination of q_j and q_k into a common thesaurus class q_{jk} produces an effective retrieval system, provided that q_{jk} is weighted in document D (or in query Q) as the sum of the weights of q_j and q_k, respectively, occurring in D (or in Q).

Proof: Consider first the difference between $E\{f(Q^*, D_1^*) \mid D_1^* \varepsilon R\}$ and $E\{f(Q, D_1) \mid D_1 \varepsilon R\}$. For the unmodified query Q and document class D_1, the contribution to the correlation of the relevant documents with the query from term q_k is zero, since Q does not contain q_k. Following the modification process, the contribution due to q_k equals r_k since there are r_k occurrences of q_k in R. Thus,

$$E\{f(Q^*, D_1^*) \mid D_1^* \varepsilon R\} = E\{f(Q, D_1) \mid D_1 \varepsilon R\} + \frac{r_k}{|R|}$$

$$= [(\sum_{i=1}^{m} r_i) + r_k]/|R|. \tag{8}$$

Similarly,

$$E\{f(Q^*, D_2^*) \mid D_2^* \varepsilon I\} = [(\sum_{i=1}^{m} \sigma_i) + \sigma_k]/|I|. \tag{9}$$

It is sufficient to show that

$$\frac{\left[\sum_{i=1}^{m} r_i + r_k\right]}{\left[\sum_{i=1}^{m} \sigma_i + \sigma_k\right]} \geq \frac{\left[\sum_{i=1}^{m} r_i\right]}{\left[\sum_{i=1}^{m} \sigma_i\right]}. \tag{10}$$

Expression (10) is true if and only if

$$r_k/\sigma_k \geq \left(\sum_{i=1}^{m} r_i\right) / \left(\sum_{i=1}^{m} \sigma_i\right) \tag{11}$$

or, equivalently, if and only if

$$r_k/(r_k + \sigma_k) \geq \left(\sum_{i=1}^{m} r_i\right) / \left(\sum_{i=1}^{m} (r_i + \sigma_i)\right). \tag{12}$$

By hypothesis, q_k and q_j are semantically related, and their accuracy is at least as large as the medium accuracy. This proves the proposition.

It is clear that when a thesaurus contains many term classes (as opposed to only one class), a repeated application of the proposition produces an effective retrieval system.

The problem with the utilization of the thesaurus process outlined earlier is the difficulty of manually building the term classes, that is, of deciding which terms are semantically related.[+] An alternative, possibly more practical, and equally effective procedure may then be suggested that does not use the concept of semantic relatedness for term grouping purposes. Specifically, two terms will be combined into a class provided they are both high-accuracy terms, and both occur in the same query.

The next proposition will demonstrate the effectiveness of the retrieval system:

Proposition 3: Let q_j and q_k be two query terms occurring in query Q, such that

$$r_\ell/(r_\ell + \sigma_\ell) \geq \left(\sum_{i=1}^{m} r_i\right) \Big/ \left(\sum_{i=1}^{m} (r_i + \sigma_i)\right) \qquad \ell = j, k.$$

If q_j and q_k are combined into a thesaurus class q_{jk}, such that the weight of q_{jk} in a document, or query, is the sum of the weights of q_j and q_k occurring in the document, or query, respectively, then the resulting retrieval system is effective.

[+] There is of course also the question of determining term accuracy. An approximation to this latter parameter may, however, be expected to become available after the system will have been in operation for some time, and appropriate relevance assessments of documents with respect to queries will have become available.

Proof: Let r_{jk} be the number of relevant documents containing both q_j and q_k. There are $(r_j - r_{jk})$ relevant documents containing q_j but not q_k. The contribution to the correlation of the original relevant documents with the original query by such documents is $(r_j - r_{jk})$. With the modified query and the modified documents, the contribution becomes $2(r_j - r_{jk})$, since the weight of q_{jk} in the query is 2. Similarly, the contribution due to the modified documents having q_k but not q_j will be $2(r_k - r_{jk})$. For those documents containing both terms, the contribution due to the original documents and query is $2r_{jk}$, but the contribution due to the modified documents and the modified query is $4r_{jk}$, since the weight of q_{jk} is 2, both in the modified query and in the modified documents.

Thus, $E\{f(Q^*, D_1^*) \mid D_1^* \varepsilon R\} =$

$$E\{f(Q, D_1) \mid D_1 \varepsilon R\} + \frac{1}{|R|}[4r_{jk} - 2r_{jk}]$$

$$+ \frac{1}{|R|}\{[2(r_j - r_{jk}) - (r_j - r_{jk})] + [2(r_k - r_{jk}) - (r_k - r_{jk})]\}$$

$$= E\{f(Q, D_1) \mid D_1 \varepsilon R\} + (r_j + r_k) / |R|. \qquad ($$

Similarly

$$E\{f(Q^*, D_2^*) \mid D_2^* \varepsilon I\} = E\{f(Q, D_2) \mid D_2 \varepsilon I\} + (\sigma_j + \sigma_k) / |I|. \qquad (1$$

It is sufficient to show that

$$\frac{\left(r_j + r_k + \sum_{i=1}^{m} r_i\right)}{\left(\sigma_j + \sigma_k + \sum_{i=1}^{m} \sigma_i\right)} \geq \frac{\left(\sum_{i=1}^{m} r_i\right)}{\left(\sum_{i=1}^{m} \sigma_i\right)}. \qquad (1$$

Expression (15) is true if and only if

$$\frac{(r_j + r_k)}{(\sigma_j + \sigma_k)} \geq \frac{(\sum_{i=1}^{m} r_i)}{(\sum_{i=1}^{m} \sigma_i)} . \qquad (16)$$

Since $\frac{(r_j + r_k)}{(\sigma_j + \sigma_k)} \geq \min \left\{ \frac{r_j}{\sigma_j} , \frac{r_k}{\sigma_k} \right\}$, expression (16) is verified by the accuracy hypothesis for q_j and q_k and the proposition is true.

Both of the suggested thesaurus (term class) transformations will therefore produce effective indexing systems.

5. Phrase Transformation

The previously described thesaurus transformation groups low-frequency (high-accuracy) terms into classes. Although each thesaurus class may be expected to have a higher document frequency (hence a lower accuracy) than the component terms, the use of thesaurus classes adds new terms with higher than medium accuracy, and the transformation proves effective. From a retrieval point of view, the use of additional content identifiers related to the original ones is a recall-improving device.

The dual, precision-improving transformation takes two or more high-frequency (low-accuracy) terms, and uses them under certain conditions to replace the original terms. This is the well-known phrase transformation. In the earlier experimental work phrases were formed of high-frequency terms that appeared in close proximity in document or query texts. [2,3] The document frequency of each phrase cannot be higher than that of any phrase component; hence the phrase accuracy is generally higher than that of the components. The replacement of two or more low-accuracy terms by one of generally higher accuracy may be expected to prove effective in retrieval. Whereas a thesaurus class was assigned a weight equal to the sum of the weights of the class elements, the phrase weight (as in the earlier discrimination value model) is taken as the average weight of the phrase components.

Proposition 4: Let q_i and q_j represent two terms included in query Q, whose accuracy is no higher than the medium accuracy of the terms in Q. If q_i and q_j are replaced by phrase q_{ij} with weight equivalent to the average weight of q_i and q_j, respectively, then the system is effective provided q_{ij} has accuracy at least as large as any of its components.

Proof: The deletion of term q_i from query Q decreases $\sum_{D_1 \varepsilon R} f(Q, D_1)$ by r_i. Similarly, the deletion of q_j reduces the sum by r_j. On the other hand, the addition of the phrase q_{ij} increases $\sum_{D_1 \varepsilon R} f(Q, D_1)$ by r_{ij}, where r_{ij} represents the number of relevant documents containing both q_i and q_j. Thus,

$$E\{f(Q^*, D_1{}^*) \mid D_1{}^* \varepsilon R\} = [\sum_{k=1}^{m} r_k - (r_i + r_j - r_{ij})] / |R|.$$

Similarly,

$$E\{f(Q^*, D_2{}^*) \mid D_2{}^* \varepsilon I\} = [\sum_{k=1}^{m} \sigma_k - (\sigma_i + \sigma_j - \sigma_{ij})] / |I|.$$

The phrase process will be effective if and only if

$$\frac{\sum_{k=1}^{m} r_k - (r_i + r_j - r_{ij})}{\sum_{k=1}^{m} \sigma_k - (\sigma_i + \sigma_j - \sigma_{ij})} \geq \frac{\sum_{k=1}^{m} r_k}{\sum_{k=1}^{m} \sigma_k} \qquad (17)$$

Expression (17) in turn is satisfied if and only if (18) is true:

$$\frac{\sum_{k=1}^{m} r_k}{\sum_{k=1}^{m} \sigma_k} \geq \frac{r_i + r_j - r_{ij}}{\sigma_i + \sigma_j - \sigma_{ij}} . \qquad (18)$$

From the phrase accuracy hypothesis, one has

$$\frac{r_{ij}}{\sigma_{ij}} \geq \max \{\frac{r_i}{\sigma_i} , \frac{r_j}{\sigma_j}\}$$

implying that $r_{ij}/\sigma_{ij} \geq (r_i + r_j) / (\sigma_i + \sigma_j)$.

Thus,

$$(r_i + r_j) / (\sigma_i + \sigma_j) \geq (r_i + r_j - r_{ij}) / (\sigma_i + \sigma_j - \sigma_{ij}). \qquad (19)$$

Since all phrase components are low-accuracy terms one has by hypothesis

$$\begin{aligned}\left(\sum_{k=1}^{m} r_k\right) \Big/ \left(\sum_{k=1}^{m} \sigma_k\right) &\geq \max \ \{r_i/\sigma_i, r_j/\sigma_j\} \\ &\geq (r_i + r_j) / (\sigma_i + \sigma_j). \qquad (20)\end{aligned}$$

Equation (18) is thus proved by combining the results of (19) and (20).

The proof is easily generalized to phrases consisting of more than two components, showing that the phrase transformation produces an effective retrieval system.

6. Conclusion

The experimental evidence obtained earlier for the discrimination value model with document collections in aerodynamics, medicine, and world affairs led to the following conclusions: [3]

a) the most effective indexing units for retrieval purposes exhibit a document frequency neither too high nor too low;

b) increases in retrieval effectiveness are produced by grouping low-frequency terms into thesaurus classes (for recall enhancement) and by using indexing phrases to replace the high-frequency components (for precision improvement).

In the present study the same transformations are shown to be formally true under the following special conditions:

a) term accuracy is assumed inversely related to the document frequency of the terms;

b) the document and query vectors are initially assumed to be binary (the earlier experimental evidence was obtained for numeric vectors using term frequency weighting);

c) the similarity measure between two vectors X and Y is assumed to be the inner product $f(X, Y) = \Sigma\, x_i\, y_i$ instead of the usual cosine measure $\cos(X, Y) = \Sigma\, x_i\, y_i \Big/ \sqrt{\Sigma\, x_i^2 \cdot \Sigma\, y_i^2}$;

d) thesaurus classes and phrases are restricted to exactly two components; in the experimental situation the size of these units is unrestricted.

None of these restrictions appears to produce fundamental difficulties in practical situations, and the formal proofs should carry over to most operational retrieval environments. Table 1 contains a summary of the effective indexing devices.

References

[1] G. Salton and M.E. Lesk, Computer Evaluation of Indexing and Text Processing, Journal of the ACM, Vol. 15, No. 1, January 1968, p. 8-36.

[2] G. Salton, C.S. Yang, and C.T. Yu, Contribution to the Theory of Indexing, Information Processing 74, North Holland Publishing Co., Amsterdam, 1974, p. 584-590.

[3] G. Salton, C.S. Yang, and C.T . Yu, A Theory of Term Importance in Automatic Text Analysis, Journal of the ASIS, Vol. 26, No. 1, January-February 1975, p. 33-44.

[4] K. Sparck Jones, A Statistical Interpretation of Term Specificity and its Application to Retrieval, Journal of Documentation, Vol. 28, No. 1, March 1972, p. 11-20.

[5] G. Salton and C.S. Yang, On the Specification of Term Values in Automatic Indexing, Journal of Documentation, Vol. 29, No. 4, December 1973, p. 351-372.

	Low-frequency terms	Medium-frequency terms	High-frequency terms
	Poor discrimination value High-accuracy	Good discrimination value Medium-accuracy	Very poor discrimination value Low-accuracy
Recall improving transform	●———	———→	Groups low-frequency terms into thesaurus classes thereby creating higher frequency units; term weights are summed
Precision improving transform	Assigns term weights in inverse document frequency order, and replaces high-frequency terms by lower frequency phrases; term weights are averaged	←———	———●

Effective Indexing Devices

Table 1