ON THE ROLE OF WORDS AND PHRASES IN
AUTOMATIC TEXT ANALYSIS

G. Salton and A. Wong *

TR 75-247

June 1975

Department of Computer Science
Cornell University
Ithaca, New York 14853

* This study was supported in part by the National Science Foundation under grant GJ 43505.

Automatic Text Analysis

G. Salton and A. Wong *

Abstract

One of the most crucial operations in automatic information retrieval is the assignment to written texts and documents of appropriate identifiers, capable of representing information content for search and retrieval purposes. This operation known as automatic indexing normally consists in assigning to the documents either single terms, or more specific entities such as phrases, or more general entities such as term classes.

A model, known as discrimination value analysis is introduced which assigns an appropriate role in the indexing operation to the terms, term phrases, and thesaurus classes. The model is used to determine effectiveness criteria for the content identifiers and to generate useful indexing policies. Experimental evidence is given to validate the theory.

1. Automatic Content Analysis

In automatic information retrieval, the first and possibly most crucial operation consists in assigning to the stored documents and to incoming user queries appropriate identifiers capable of representing information content. The difficulty of the task is illustrated by the fact that in practically all

existing, operational retrieval systems this indexing operation is carried out manually by trained indexers or subject experts, rather than automatically.

Quite a few observers in fact claim that the reduction of written texts to individual "units of expression" for the representation of content is inherently impossible, and that retrieval, or text processing systems that are based on such content identifiers can never operate satisfactorily:

"It is a logical category mistake to assume that a word, or a phrase, contains information in the same sense in which a statement does. In spite of *prima facie* appeal, the information content of a statement is not the sum, or combination, of the information content of its constituent phrases ..." [1]

This type of argument leads to the conclusion that index terms and phrases are not substitutes for more complete content identifications, and that term sets cannot therefore function as miniature or condensed documents. The latter assertion is reinforced by showing that the juxtaposition of terms in the language is in no way comparable to the intersection, or union, of the document sets identified by the corresponding terms. Thus, the set of documents related to "fish food" is not identical with the set identified by both "fish" and "food", and neither is a "Newfoundland dog" something characterizable by the term pair "Newfoundland" and "dog". [1]

Instead of reducing the representation of written texts to a set of simple terms only, a full theory of language understanding appears to be needed which would account for the complete stated and implied content of the texts. Such a theory of language understanding should be capable of identifying not only an

*Computer Science Department, Cornell University, Ithaca, NY, 14853

This study was supported in part by the National Science Foundation under grant CJ 43505.

appropriate set of content indicators, but also two main types of relations between indicators: [2]

- a) the logical-semantic relations between text units which are dependent on the world knowledge, and the social context within which a given area of discourse is placed;
- b) the linguistic-semantic relations which are dependent on the linguistic context and are derivable from a knowledge of the "deep" structure of the texts.

To determine the logical-semantic relations an encyclopedia, or semantic net, is often thought to be required which would identify the scope and extent of a given subject area, and the known relationships between the concepts included in the field. The linguistic-semantic relations, on the other hand, are obtainable by using a combined syntactic-semantic analysis to generate a detailed structure of the written texts. When the world (encyclopedic) knowledge is combined with the linguistic analysis, a text should then be representable as a series of "inference chains" representing the line of thought expressed in the text, including both the stated as well as the unstated assumptions and conditions. [3]

Unfortunately, a large number of unresolved problems interfere at present with the utilization of language understanding systems. There is uncertainty about the primitive entities needed to represent information content, and about the relations which may exist between individual concepts. Furthermore, it is unclear how best to represent the dictionary, encyclopedia, semantic map, etc., which may be used to specify the necessary world knowledge. Finally, no viable prescriptions exist for resolving many ambiguities, discarding useless context, and generally reaching the proper conclusions in the construction of the inference chains.

Moreover, even if a viable theory of language understanding could in fact be implemented, it would constitute only a beginning, and not the end for information retrieval purposes. Additional problems would then arise in the determination of appropriate search strategies through the respective inference chains, and in deciding on retrieval, or nonretrieval, of the various documents based on available user relevance specifications.

In practice, it is then necessary for the foreseeable future to fall back on less sophisticated methodologies. While a language understanding theory may not be available, a good deal is known about the structure of written texts: certain words are used frequently, others are rare; some words are quite short, others are much longer; some words occur in every document in a given collection, others are concentrated in only a few documents.

Based on such objective criteria, a theory of term value is developed in the next section which is later incorporated into an appropriate indexing strategy.

2. Term Value and Indexing Theories

A great many studies have been made in the recent past of the occurrence characteristics of words in running texts, and of the manner in which single terms can be combined with each other to form larger text units such as phrases. It is not possible in the present context to review in detail all the known facts in this connection. It may suffice to mention only the most important properties first for single terms and then for term combinations.

Single terms may first be characterized by their overall frequency of occurrence in written texts. The following properties may be cited in this connection: [4]

- a) High-frequency words that occur repeatedly in written texts are shorter on the average (they include fewer phonemes) than low-frequency words; they often exhibit a large number of different meanings; and they enter as components into many different permutations (phrases).
- b) Low-frequency words, on the other hand, are often longer in terms of phonemes; they exhibit fewer distinct meanings, and they are used for a smaller number of different phrases.

Few conclusions can be drawn about the value of a term in retrieval based on raw word frequency statistics alone. Experience indicates, however, that very high-frequency words are not optimal as content indicators because such words cannot be used to discriminate the texts from each other. Very low-frequency terms tend to be more specific, but since they occur so rarely, they lead to few matches between incoming queries and stored documents and their effectiveness for retrieval is therefore also questionable.

To learn more about the usefulness of single terms, it is necessary to look at the frequency distribution of the terms across the documents of a collection, and at the total document frequency of a term (the number of distinct documents in which the term occurs, or to which it is assigned). Once again, extreme distributions are not optimal:

- a) Terms whose frequency distribution is quite flat, in that they occur with approximately equal frequency in many documents of a collection, are not very effective as content indicators because they cannot distinguish the documents in which they occur from each other.

- b) Terms whose frequency distribution is very concentrated and peaked — they occur with high-frequency in one or two documents only and are absent from the rest — are also not optimal because their inclusion as part of a user query affects very few items.

The best evidence obtainable from a frequency characterization of word occurrences in texts indicates that the most useful content indicators are terms whose overall occurrence frequencies are neither too large nor too small, and whose frequency distributions across the documents of a collection are skewed — they occur with high-frequencies in some documents and with much lower frequencies in others.

One additional complicating factor must be mentioned in connection with the analysis of term importance based on occurrence frequencies and frequency distribution characteristics. Obviously, among all the terms possessing a given frequency profile, the most valuable will be those that are concentrated in documents specified by the users as relevant to their respective queries, rather than those prevalent mostly among the nonrelevant items. This leads to the concept of term precision as the balance between the relevant and the nonrelevant documents in which a term occurs. [5] Ideally, the term precision should be used as a term value indication in addition to the frequency characteristics mentioned earlier. In practice, the term precision may not be easy to utilize in the absence of a sufficient number of reliable relevance assessments of documents with respect to user queries.

While formal frequency properties may be adequate to characterize the single terms, additional criteria are normally introduced for larger text units, such as phrases and term classes. The aim in using either phrases as content identifiers, or complete classes of related or synonymous terms is the same:

namely the assignment of entities which are more effective for content representation than single terms. The result achieved is however different in the two cases: Phrases are used when it becomes necessary to generate narrower, more specific content identifiers than the component single terms; the effect is the replacement of broad, high-frequency terms by lower frequency phrase units, thereby enhancing the precision of the search. The assignment of a complete term class, on the other hand, implies the addition of new terms not originally present; this replaces a single, often low-frequency, term by a whole class of terms whose occurrence frequency is then necessarily higher than that of any class member. The use of term classes instead of single terms normally produces better recall.*

The methodologies used in finding either phrases or term classes are somewhat parallel. In either case term relationships are specified by using one or more of the following criteria: [6]

- a) positional: using the location in the given paragraph structure in which a given term occurs; special provisions are often made for terms occurring in titles, footnotes, bibliographies, etc.
- b) semantic: using semantic properties of terms such as synonymy; synonymous term classes are often prespecified in a thesaurus or semantic map.

*Precision is the proportion of retrieved material that is relevant, whereas recall is the proportion of relevant material that is retrieved. High precision implies the rejection of most extraneous materials; high recall, on the other hand, leads to the retrieval of most items that are considered relevant.

- c) pragmatic: incorporating the social context within which the terms are used, including for example the size, or weight of a concept, and the environment in which the concept is likely to arise.
- d) statistical: taking into account the tendency for two or more terms to cooccur in the same sentences, or the same documents of a collection; term-term association measures are used as a basis both for automatic phrase recognition, and for the construction of automatic term classifications and thesauruses.
- e) nonstatistical: using, for example, the presence of special "cue words" to determine term relations, or constructing a logical paraphrase of each sentence which exhibits the logical relations between sentence units, or using structural and syntactic relationships.

Some of the criteria are much more easily utilized than others — for example, positional characteristics are often straightforward to use, whereas the semantic properties may be impossible to incorporate effectively. In practice, statistical criteria are often used, particularly for the determination of relations that are directly specified in the text; these may be supplemented by simple syntactic methods that are relatively inexpensive to implement.

Baxendale, for example, argues that noun phrases and noun-adjective groupings constitute the best "units of expression". Accordingly, she detects simple syntactic entities such as punctuation marks, prepositions, conjunctions, etc., and selects as phrases expressions that are bracketted between such special entities. [7] Hillman goes a step further by attempting to determine for each sentence a logical paraphrase in which the individual assertions are properly specified and related to each other. Rules are then used to specify as a phrase, those text units which are found in "referential position" within the sentences of a document. [8]

Many other procedures have been advocated for the determination of effective content indicators, often using methods constructed specifically for a given environment, or a given subject area. Some of these processes are quite valuable within their contexts, but none of them can be said to operate perfectly, and none provides a complete framework in which the various pieces are properly motivated and accounted for.

A framework which assigns a specific role in content analysis to single terms, term phrases, and term classes is presented in the next section in the form of the discrimination value model. This model is then evaluated in the remaining portions of this study.

3. Discrimination Value Model

The discrimination value model is one of several possible systems capable of rating the potential index terms in accordance with their usefulness as document discriminators; the model also offers the advantage of providing a reasonable physical interpretation for the indexing process. [9] Consider in particular, a collection of documents each identified by a set of content identifiers, or index terms. The index term sets for two given documents can be compared to produce a similarity coefficient measuring the closeness between the respective documents. The existence of the term sets representing the various documents, and the possibility of computing similarity measures between documents can be used to define a document space for the collection. In such a space two documents appear in close proximity when their similarity coefficient is large; contrariwise, documents exhibiting little similarity are widely separated in the document space.

Since a one-to-one relation exists between a set of index terms used to characterize a collection of documents and the space formed by the indexed documents, the problem of choosing an effective term set may be transformed into one of generating a useful document space. In this connection one may conjecture that a document space which is "bunched up", in the sense that all documents exhibit somewhat similar term sets, or "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 may provide an ideal retrieval situation since some documents may then be retrieved while others can be rejected.

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 document of a collection. Specifically, a good discriminator is one which, when assigned as an index term to the documents of a collection, will render the documents less similar to each other; that is, its assignment decreases the space density. A poor discriminator, on the other hand, increases the space density when assigned.

For each potential index term, a discrimination value (DV) can then be computed as the difference between the space densities before and after its assignment. The greater the difference in space densities, the more the space will have spread and therefore the better the term will function as a discriminator. The space density may be defined as the average pairwise similarity between all documents, or more efficiently as the sum of the similarity coefficients between each document and a dummy "centroid" document defining the center of the space.

The best discriminators will generally exhibit positive discrimination values (the space spreads when they are assigned); correspondingly, the worst

discriminators have negative discrimination values. In the middle, one finds a large number of terms 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 highest discrimination value corresponds to the lowest discrimination rank.

It may be of interest to inquire into the relationship between the discrimination value of a term and some of the frequency parameters introduced earlier. The following conclusions are reached from a study of the indexing vocabularies in several different subject areas relating the document frequency (DF) of a term to its discrimination value: [10]

- a) terms with very low document frequencies (DF) that may be assigned to only one or two documents in a collection are generally poor discriminators, with discrimination ranks in excess of $t/2$ for a total of t existing terms;
- b) terms with high document frequencies, comprising those that are assigned to more than 10 percent of the documents in a collection are the worst discriminators, with average discrimination ranks near t ;
- c) the best discriminators are those whose document frequency 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.

These relationships are studied in more detail in the next section.

4. Single Term Analysis

To study the properties of various indexing vocabularies, three document collections are used in different subject areas including aerodynamics, medicine, and world affairs. The CRAN 424 collection consists of 424 documents in aerodynamics, including 2651 distinct terms. The MED 450 collection comprises 450 documents in medicine with a total of 4726 terms. And the TIME 425 collection includes 425 articles in the field of world affairs published in Time magazine during 1963; the total number of terms for the TIME collection is 14,098.

The vocabulary is fairly homogeneous for CRAN and MED, and much more heterogeneous for TIME, as indicated by the large number of distinct terms for that collection. The average document frequency is approximately equal for CRAN and TIME, each term being assigned on the average to 10 different documents; for MED, the average document frequency per term is only about 5.

Fig. 1 relates the frequency and discrimination value parameters for 30 distinct terms included in the TIME collection, consisting of sequences of ten terms with discrimination ranks between 1 and 10, between 3780 and 3789, and finally between 7560 and 7569, respectively. The following parameter values are given for the terms in these DR ranges:

- a) the average document frequency (DF), that is, the average number of documents to which these ten terms are assigned;
- b) the average term frequency (TF), that is, the average total number of occurrences in the collection for these ten terms;
- c) the average discrimination value (DV) for the ten terms;
- d) the average document frequency distribution for the term sets.

From the small table on the right-hand side of the graph it is obvious that the best discriminators (average DV value 0.1130) are those terms whose document frequency averages 21 out of 425; these terms occur about 130 times in all on the average in the documents of the collection. Next in decreasing DV order are the low-frequency terms whose average discrimination value is near zero. These terms have approximately equal document and term frequencies of about 10, indicating that each term occurs once in each of 10 documents. Finally, the worst discriminators with DR between 7560 and 7569 have a negative average discrimination value of -0.9295; their average document frequency is about 188 — that is, each term is assigned to nearly 188 documents out of 425 — and their average total frequency is 476. The experimental parameter values for these 30 terms thus confirm the correctness of the earlier generalization that the best discriminators are terms with average rather than extreme frequency parameters.

Consider now the graphs of Fig. 1. The low-frequency terms have document frequencies restricted to 1 or 2. The high-frequency terms with average DR of 7565 occur once in over 86 documents, twice in 46 more documents, three times in about 21 documents, four times in over 12 documents, and so on. It is not surprising that such ubiquitous terms cannot be used to discriminate among the documents of the collection. The best terms with DR's between 1 and 10 occur with low-frequency in only a few documents, with higher frequencies in some additional documents, and not at all in the remainder.

The various relationships are summarized in Table 1 for the three collections under examination. This table gives the DV ranges for the term sets pertaining to the three subject areas. Also included are the discrimination ranks for the large number of terms with near zero discrimination values and for the relatively small number of high-frequency terms with negative discrimination values. The best terms with positive discrimination values have document frequencies between 10 and 49, 9 and 14, and 11 and 29 for the CRAN, MED, and TIME collections, respectively.

Table 2 gives actual terms occurring in certain ranges of the discrimination rank (DR) order. Going from left to right, one first meets quite specific terms with low DR values; then as the DR ranks increase the terms become less specific but also less frequent in the collection. Finally, on the right hand side of the Table appear the most general, high-frequency terms.

One may ask whether these statistics can be used to decide on a good indexing policy. The first possibility is to restrict the indexing vocabulary to only those terms which exhibit the preferred frequency characteristics. This could be done by eliminating the very high and very low-frequency terms. The evidence available from a variety of term deletion experiments indicates that such a strategy is not very effective for retrieval purposes. [11] The problem is that the removal of the high-frequency terms tends to depress the recall, whereas deleting the low-frequency indicators affects the precision. For an average query the reduced vocabulary performs less well than the full term set.

If the elimination of the poor discriminators does not produce the desired result, an alternative strategy might consist in transforming those terms

exhibiting inadequate discrimination characteristics into better terms with more favorable parameter values. The following indexing strategy suggests itself [10]:

- a) Terms whose document frequency lies between $n/100$ and $n/10$ for n documents should be used for indexing purposes directly; these terms include the vast majority of the good discriminators.
- b) Terms whose document frequency is too high comprise the worst discriminators. These terms are too general, or too broad, and their use produces unacceptable precision losses. These terms should be transformed into lower frequency units by using them as components of appropriate indexing phrases.
- c) Terms whose document frequency is too low are so rare, or specific, that they cannot retrieve an acceptable proportion of the relevant documents; their use depresses the recall performance. Such terms should be transformed into higher frequency units by including them in classes of related terms, specified for example by a thesaurus.

Such a policy then accounts for three different entities useful for document content representation: single entities are used for terms whose discrimination value is already acceptable; term phrases are constructed from high-frequency components with negative discrimination values; and term classes are generated of the low-frequency components with near-zero discrimination values.

The phrase and thesaurus class characteristics are examined in the next two sections.

5. Analysis of Phrases

In order to evaluate the use of phrases for content identification, a number of different phrase generation procedures should ideally be tested.

Linguistic techniques might be given special attention in this respect since these are sometimes believed to be essential for the formation of useful phrases. For present purposes, a simpler strategy is adopted which substitutes term cooccurrence characteristics for an expensive syntactic analysis. In so doing, the hope is that if the simple process proves effective, even better results might be obtained with more sophisticated phrase formation methods.

The following phrase generation process is adopted experimentally [10]:

- a) phrases are defined by using query texts;
- b) common function words are removed and a suffix deletion method is used to reduce the remaining query words to word stems;
- c) the remaining word stems are taken in pairs, and each pair defines a phrase, provided that the distance in the text between the two phrase components does not exceed two (at most one intervening word may occur between components);
- d) phrases for which both components are identical are eliminated;
- e) duplicate phrases, where all components match an already existing phrase are eliminated;
- f) a weight is assigned to each phrase as a function of the weights of the individual phrase components.

In this process, a phrase is defined as a pair of cooccurring terms in a query sentence, with an added restriction on the distance permitted between the two terms.

Twenty four user queries were available with each of the three collections for phrase generation purposes. The number of phrases generated by the process

ranged from 267 for the CRAN collection to 379 for MED. Each phrase was placed into one of six frequency classes depending on the document frequency of the term components. Three frequency classes were used for individual terms, including low (L), medium (M), and high (H) frequency terms. The phrase classes which result thus range from LL (low-low) where two low-frequency terms are combined, to HH (high-high) for phrases consisting of two high-frequency terms.

The document frequencies used to define the term frequency classes for the three collections are shown in Table 3; the number of resulting phrases is given for the various frequency classes in Table 4. It may be noticed that the vast majority of the phrases falls into either the M- or the H-phrase groups (92 percent and 93 percent for CRAN and Time, respectively). The relatively large number of low-frequency terms included in the medical collection do produce about 90 L-group phrases for MED.

The actual phrase formation process is illustrated in the example of Table 5 for a TIME query in world affairs. Each phrase is listed in its frequency class together with the document frequency of the phrase and the phrase precision. The latter is defined as the number of phrase occurrences in documents defined as relevant by the user population divided by the document frequency of the phrase. Phrase precision may be an indication of phrase usefulness for retrieval purposes.

Average phrase precision figures are given for the three collections in Table 6 for phrases in five of the six frequency classes. It may be noticed that the large number of HH phrases exhibits rather low precision,

whereas the phrases which combine one low-frequency component with one of higher frequency (LH and LM, in particular) seem much more effective as a group. This appears to confirm some early observations by Baxendale [7] in which she suggests that the combination of one high-frequency component (such as "system", "measurement", or "theory") with a low-frequency modifier could improve discrimination beyond that of either term individually.

The usefulness of the phrase process for document discrimination purposes is reflected in the output of Table 7. In that Table only those query phrases are taken into account which also occur in the documents of the respective collections. The total number of phrases listed is thus lower than that given earlier in Tables 4 and 6. The evaluation criterion shown in Table 7 is the difference in discrimination ranks between the phrases on the one hand, and the corresponding phrase components on the other. (Obviously, when phrases are used for content identification, a phrase discrimination value can be computed as the difference in document space densities with and without the phrase assignment).

It may be seen from Table 7 that for all three collections the average discrimination rank of the phrases is substantially lower than for the terms which are used as phrase components. This is also reflected by the actual number of phrases producing improvement in DR ranks which lies between 60 and 70 percent for the three collections.

A complete retrieval performance evaluation is given for the three collections in Tables 8, 9, and 10. In each case, average precision values are given for the 24 user queries at ten specific recall points, varying from

0.1 to 1.0 in steps of 0.1.* The Tables are separated into four parts reflecting the performance of the various phrase groups, including H-group, M-group, M- and H-group, and All phrase groups, respectively (the phrase frequency groups are defined in Table 4). Included in part a) of the Tables is a control run in which an automatic word stem procedure producing single terms only is used for indexing purposes. All other runs are based on the addition of phrases with or without deletion of some of the phrase components. Since the discrimination value theory indicates that poor high-frequency terms should be replaced by better lower-frequency phrases, two types of term deletion methods are used:

- a) deletion of all high-frequency (H) terms following the addition of phrases, whether or not the deleted terms actually function as phrase components;
- b) deletion of certain terms (either H- or M-terms) only for those terms actually used as phrase components.

The best performance corresponding to each of the subsections of Table 8 to 10 is identified by a vertical bar adjacent to the corresponding precision value. The best performance for any of the phrase modification methods is marked with a double bar.

A number of conclusions are immediately obvious:

- a) the addition of phrases produces considerable performance improvements over the use of single terms alone;

*In an evaluation based on recall-precision output one assumes that the preferred output produces the highest precision for a given level of recall; this implies that the proportion of nonrelevant items included among the retrieved ones is minimized.

- b) the phrase process appears to be more powerful than the use of thesaurus classes in addition to single terms, shown for comparison purposes in section d) of the Tables;
- c) in practically all cases, the use of phrases is better with a corresponding deletion of the single terms than without; this confirms the behavior predicted by the discrimination value theory which indicates that poor discriminators can be transformed into more effective terms by phrase and thesaurus frequency transformations;
- d) the use of both M- and H-group phrases is generally more effective than the use of either M- or H-group phrases alone; this confirms the notion that phrases consisting of the juxtaposition of one high- and one lower frequency component may be most effective;
- e) the best overall phrase methodology consists of the addition of M- and H-group phrases, or of All phrases, followed by deletion of the high-frequency (H) terms.

Some minor performance differences are noticeable for the three document collections. For CRAN, the H-phrases alone are not very effective, although these phrases perform well with the other two collections. The CRAN collection is also distinguishable from the other two by the fact that the use of All phrases is marginally better than the combined use of both M- and H-phrases; furthermore, term deletion should not occur for the CRAN collection at the very high recall end of the performance spectrum. The performances are reasonably parallel for MED and TIME, with the exception that for the TIME collection the M-group of phrases is best used without term deletion.

A summary of the best phrase performances appears in Table 11 for the three collections.

6. Analysis of Term Classes

The discrimination value theory predicts that a grouping of low-frequency terms into higher frequency term classes should be useful to enhance the recall performance. In conventional retrieval environments, term classes are often specified by a thesaurus, and great care is taken to insure that terms entered into a common thesaurus class be closely synonymous. In a low-frequency thesaurus, a looser grouping of terms may be permissible, since low-frequency terms will generally exhibit few different meanings.

A number of obvious methodologies suggest themselves for an automatic generation of low-frequency term classifications (thesauruses):

- a) existing thesauruses for various subject areas might be transformed into low-frequency thesauruses by eliminating high-frequency terms, or by eliminating all classes whose frequency exceeds a given threshold;
- b) each term class might be generated so as to include the complete set of low-frequency terms occurring in a given document;
- c) a given term class might be created for all low-frequency terms exhibiting the same word stem;
- d) in areas such as world affairs, one class might represent all low-frequency terms falling into a given, broad subject area (for example, one class per country or geographic subdivision for the Time magazine material).

For current experimental purposes strategy a) is chosen by utilizing existing thesauruses in the three subject fields and applying frequency restrictions. Thus, all thesaurus classes whose document frequency exceeded 60, 40, and 60 for CRAN, MED and TIME, respectively were eliminated.

The resulting thesaurus class characteristics are shown in Table 12. It may be noted from the Table that relatively limited term classifications are produced. A large number of classes are obtained only for the MED collection. The discrimination value theory is confirmed for the thesaurus transformation by the output of Table 12 as it was earlier for the phrases. Indeed almost all of the thesaurus classes exhibit higher discrimination ranks than the corresponding terms included in the classes. Substantial average discrimination rank differences (DRD) are produced for the CRAN and TIME collections; a small loss occurs for the MED thesaurus classes, probably due to the fact that the MED thesaurus consists of a large number of very small classes.

Consider now the problem of distinguishing useful thesaurus classes from those that are less effective, or, alternatively, of predicting thesaurus class performance. Table 13 exhibits a number of typical thesaurus classes for the three collections, including some with improved discrimination rank, and others with reduced ranks. The three classes shown in the top half of the Table which lead to DR improvements are produced by a grouping of several low-frequency (L) terms. The bottom classes, on the other hand, are either very small, consisting of one or two medium-frequency terms, or else they include a high-frequency (H) term whose large occurrence frequency reduces the importance of all other terms in the same class. The output of Table 13 is typical of the complete thesauruses constructed for the current experiments, and leads to the following class characterizations:

- a) the most effective thesaurus categories (term classes) are those which include a large number of low-frequency (L) terms;
- b) the least effective classes either consist of only one or two terms, or else they include terms with unequal frequency characteristics permitting the high-frequency terms to overcome the others in the same group.

The recall-precision output of Table 14 shows that the low-frequency thesaurus transformation brings performance improvements of about ten percent, compared with the single term word stem process. For the TIME collection the improvement is smaller. This is explained by the very specific fact-retrieval-type user queries available with this collection; a recall-enhancing transformation will not help the performance of narrow precision-oriented queries. The output of Table 14 also shows that, in general the complete set of thesaurus classes should be used for retrieval purposes rather than only those classes with improved discrimination ranks.

A final question to be resolved concerns the effectiveness of the total discrimination value model, including the single terms with adequate discrimination values, the high-frequency nondiscriminators transformed into phrases, and the low-frequency nondiscriminators.

The following strategy is used for this purpose:

- a) all thesaurus classes (see Table 12) are used, and terms included in the thesaurus are not deleted from the single term list;
- b) all phrases (see Table 7) are used, but high-frequency phrase components are deleted whose document frequency exceeds 45, 18, and 49 for CRAN, MED, and TIME, respectively.

The corresponding recall-precision results are shown in Table 15 including both single-term word stem results, phrases with high-frequency component deletion, and thesaurus class transformation. A double bar is used to denote the best overall recall-precision performance from among those exhibited in the present study. It is clear from Table 15 that the combined strategy using single terms plus phrase and thesaurus transformations outperforms all other content analysis systems. For the fact-retrieval type questions used with the TIME collection (where high precision is at a premium), the phrases alone perform best at the low recall end of the spectrum. A few isolated precision points exist for all three collections where one or another of the phrase strategies of Tables 8, 9 and 10 is optimal.

It is believed that the experimental evidence included in this study confirms the premises of the discrimination value theory:

- that the best discriminators are terms of medium document frequency;
- that phrases should be formed from high-frequency terms for improved precision performance;
- that term classes should be formed from low-frequency terms for improved recall performance;
- that the best term classes are those which group a number of low-frequency terms to form a thesaurus category;
- that the best phrases are built by juxtaposing one high-frequency component with one or more components of lesser frequency.

It is hoped that such a strategy can be tried out in practice under operational conditions.

References

- [1] Y. Bar Hillel, Language and Information, Addison Wesley Publishing Co., Reading, Mass., 1964, chapter 18.
- [2] C.A. Montgomery, Linguistics and Information Science, Journal of the ASIS, Vol. 23, No. 2, May-June 1972, p. 195-219.
- [3] R.C. Schank and K.M. Colby, Computer Models of Thought and Language, W.H. Freeman and Co., San Francisco, 1973.
- [4] G.K. Zipf, Human Behavior and the Principle of Least Effort, Addison Wesley Publishing Co., Cambridge, Mass., 1949.
- [5] C.T. Yu and G. Salton, Precision Weighting — An Effective Automatic Indexing Method, to appear in Journal of the ACM, also Technical Report 75-232, Department of Computer Science, Cornell University, April 1975.
- [6] H.P. Edmundson and R.E. Wyllys, Automatic Abstracting and Indexing — Survey and Recommendations, Communications of the ACM, Vol. 4, No. 5, May 1961, p. 226-234.
- [7] P.B. Baxendale, An Empirical Model for Machine Indexing, Machine Indexing, Papers Presented at the Third Institute on Information Storage and Retrieval (13-17 Feb. 1961), Center for Technology and Administration, American University, p. 207-218.
- [8] D.J. Hillman, Negotiations of Inquiries in an On-Line Retrieval System, Information Storage and Retrieval, Vol. 4, 1968, p. 219-238.
- [9] G. Salton, A Theory of Indexing, Regional Conference Series in Applied Mathematics No. 18, Society for Industrial and Applied Mathematics, Philadelphia, 1975.
- [10] 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.
- [11] 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.

Number of Documents
With Corresponding
Frequency

Average Document Frequency Distribution
for 10 Terms in Various Ranges
of Discrimination Value (Time 425 Collection)

Fig. 1

Term Characteristics	CRAN 424	MED 450	TIME 425
Total number of terms	2651	4726	7569
Discrimination value DV	$-2.94 \leq DV \leq 0.26$	$-5.02 \leq DV \leq 0.13$	$-1.86 \leq DV \leq 0.25$
Document frequency DF for BEST discriminators	$9 < DF < 50$	$8 < DF < 15$	$10 < DF < 30$
Discrimination ranks (DR) for POOR low-frequency terms	$0 \leq DV < .00732$ 560 - 2577 (2518 terms or 95%)	$0 \leq DV < .01046$ 500 - 4585 (4086 terms or 85%)	$0 \leq DV < .00150$ 2240 - 7163 (4924 terms or 70%)
Discrimination ranks (DR) for VERY POOR high-frequency terms	$DV < 0$ 2578 - 2651 (73 terms or 3%)	$DV < 0$ 4586 - 4726 (140 terms or 3%)	$DV < 0$ 7164 - 7569 (405 terms or 6%)

Discrimination Value Ranks for Various Frequency Ranges

Table 1

	Collection Specific Terms	Less Specific Terms (rarer)	Very Rare Terms	Bad Discrimi- nators (frequent)
CRAN 424	DR 1 - 50	DR 1000 - 1050	DR 2000 - 2050	DR 2566+
	panel nozzle transit wing reynold gas	step accept appreciable shock-tube mathematical	bandwidth single-element turbo-machine unbanked country	line hypersonic type aerodynamic simple
MED 450	DR 1 - 50	DR 2000 - 2050	DR 4000 - 4050	DR 4586+
	marrow amyloidosis hepatitis mycoplasma coronary	mobilisation adjoining murmur diastole phenylpropionate	high-pressure reemphasize straight intercellular iodine-containing	distinct single component improve adult
TIME 425	DR 1 - 50	DR 2000 - 2050	DR 6000 - 6050	DR 7164+
	buddhist Diem Karamanlis Geneva Senghor Nato	prevent ring Kristallnacht Souphanouvong mourning toll	Bach Mohamed Elkassem Yvette unmatched	safe paper taking son girl blood

Typical Terms in Decreasing Discrimination

Rank (DR) Order

Table 2

Type of Term	CRAN 424	MED 450	TIME 425
H: high-frequency	$DF \geq 50$	$DF \geq 15$	$DF \geq 30$
M: medium-frequency	$10 \leq DF \leq 49$	$9 \leq DF \leq 14$	$11 \leq DF \leq 29$
L: low-frequency	$DF \leq 9$	$DF \leq 8$	$DF \leq 10$

Document Frequency (DF) of Terms Used
to Classify Terms as H, M, L
Table 3

Frequency Group	Phrase Category	CRAN 424	MED 450	TIME 425
L-group	LL LM	3 } 11% 25 }	41 } 24% 50 }	2 } 7% 15 }
M-group	LH MM MH	23 } 66% 33 } 114 }	90 } 54% 24 } 89 }	23 } 50% 23 } 110 }
H-group	HH	69 } 26%	83 } 22%	142 } 43%
Total	LL to HH	267 } 100%	379 } 100%	315 } 100%

Number of Distinct Phrases Formed
from Queries in Various Frequency Classes
Table 4

Original query

Kennedy administration pressure on Ngo Dinh Diem to stop suppressing the Buddhists

Phrase Type	Phrase	Doc. Freq. of Phrase	Phrase Precision
HH	Kennedy pressure	20	2/20
MH	Kennedy administration	9	0/9
	administration pressure	4	0/4
	pressure Ngo	9	4/9
	pressure Dinh	9	4/9
	Dinh stop	4	1/4
	Diem stop	4	1/4
	stop Buddhist	4	1/4
MM	administration Ngo	1	0/1
	Ngo Dinh	20	7/20
	Ngo Diem	20	7/20
	Dinh Diem	20	7/20
LM	Diem suppression	1	0/1
	Suppression Buddhist	1	0/1

$$\text{Phrase precision} = \frac{\text{No. of phrase occurrences in relevant documents}}{\text{Document frequency of phrase}}$$

Typical Phrase Formation Process

Table 5

Type of Phrase	HH	MH	LH	LM	MM
CRAN 424	69 phrases 0.1630	114 phrases 0.2280	23 phrases 0.4383	25 phrases 0.5554	33 phrases 0.1973
MED 450	83 phrases 0.2403	89 phrases 0.3978	90 phrases 0.4787	50 phrases 0.7078	24 phrases 0.6351
TIME 425	142 phrases 0.2012	110 phrases 0.4659	23 phrases 0.4181	15 phrases 0.6000	23 phrases 0.5507

Average Phrase Precision for Various Types
of Phrases

(LL phrases omitted)

Table 6

L	low document frequency		Number of Phrase Occurrences in Relevant Documents
M	medium document frequency	Phrase Precision =	
H	high document frequency		Document Frequency of Phrase

	CRAN 424	MED 450	TIME 425
Number of distinct phrases formed from queries which also occur in documents	252	234	315
Number of distinct phrases producing improvement in discrimination rank (DR)	152	159	190
Number of distinct phrases not producing improvement in DR rank	100	75	125
Average rank in discrimina- tion value order (DR) of			
phrases	1151.8	1704.9	3562.0
terms entering into phrases	1483.0	2632.4	4293.7
difference	-331.2	-927.5	-721.7

Phrase Characteristics and Performance

Table 7

Recall	Standard Word Stem (single terms only)	Addition of H-group phrases	Addition of H-group phrases plus deletion of H terms	Addition of H-group phrases, deletion of H phrase components
.1	.6844	.6209	.6377	.6662
.2	.5303	.4837	.6117	.5797
.3	.4689	.4739	.5314	.5192
.4	.3482	.3263	.4474	.4106
.5	.3131	.3134	.3587	.3503
.6	.2556	.2729	.3030	.3064
.7	.1989	.1931	.2631	.2590
.8	.1631	.1610	.1916	.1816
.9	.1265	.1347	.1432	.1244
1.0	.1176	.1291	.1273	.1144

a) Performance with H-group phrases

Recall	Addition of M-group phrases	Addition of M-group phrases plus deletion of H terms	Addition of M-group phrases, deletion of M phrase components
.1	.7759	.8385	.8264
.2	.6095	.7105	.6393
.3	.5284	.5629	.5234
.4	.4417	.5164	.4153
.5	.3898	.4119	.3365
.6	.3221	.3409	.2171
.7	.2415	.2542	.1794
.8	.1837	.1468	.1203
.9	.1369	.0990	.0731
1.0	.1254	.0990	.0726

b) Performance with M-group phrases

Phrase Performance for CRAN 424 Collection
(average precision at ten recall levels)

Table 8

Recall	Addition of M- and H-group phrases	Addition of M- and H-group phrases and deletion of H-terms	Addition of M- and H-group phrases and deletion of H-phrase components
.1	.7643	.8317	.8740
.2	.6089	.7208	.8865
.3	.5236	.5916	.5606
.4	.4208	.5068	.4699
.5	.3830	.4045	.3936
.6	.3268	.3556	.3402
.7	.2312	.2881	.2875
.8	.1842	.1807	.1777
.9	.1399	.1317	.1259
1.0	.1317	.1175	.1124

c) Performance of M- and H-group phrases

Recall	Addition of all phrases	Addition of all phrases and deletion of H- terms	For comparison Thesaurus added (all classes)
.1	.7712	.8155	.7269
.2	.6077	.7248	.5833
.3	.5615	.5836	.5192
.4	.4242	.5099	.4219
.5	.3854	.4107	.3613
.6	.3573	.3576	.2867
.7	.2685	.2986	.2296
.8	.2025	.1914	.1756
.9	.1524	.1417	.1289
1.0	.1439	.1272	.1131

d) Performance of All Phrases

Phrase Performance for CRAN 424 Collection
(average precision at ten recall levels)

Table 8 (cont.)

Recall	Standard Word Stem (single terms)	Addition of H-group phrases	Addition of H-group phrases plus deletion of H-terms	Addition of H-group phrases, deletion of H-phrase components
.1	.7891	.7870	.8593	.8521
.2	.6750	.7133	.7955	.7760
.3	.5481	.6095	.6841	.6515
.4	.4807	.5301	.6389	.5889
.5	.4384	.4811	.5829	.5443
.6	.3721	.4006	.5227	.4743
.7	.3359	.3545	.4729	.4251
.8	.2195	.2371	.3210	.2925
.9	.1768	.1775	.2543	.2263
1.0	.1230	.1327	.1932	.1605

a) performance with H-group phrases

Recall	Addition of M-group phrases	Addition of M-group phrases plus deletion of H-terms	Addition of M-group phrases, deletion of M-phrase components
.1	.8483	.8475	.7193
.2	.7233	.7237	.5898
.3	.6241	.6297	.4747
.4	.5464	.5639	.4410
.5	.4872	.5162	.3586
.6	.4097	.4574	.2947
.7	.3548	.4296	.2562
.8	.2579	.2968	.1639
.9	.2059	.2470	.1524
1.0	.1313	.1467	.0629

b) Performance with M-group phrases

Phrase Performance for MED 450 Collection
(average precision at ten recall levels)

Table 9

Recall	Addition of M- and H-group phrases	Addition of M- and H-group phrases and deletion of H-terms	Addition of M- and H-group phrases and deletion of H-phrase components
.1	.8108	.8919	.8853
.2	.7531	.8126	.8089
.3	.6770	.7332	.7122
.4	.5651	.6570	.6171
.5	.5117	.5952	.5546
.6	.4321	.5206	.4695
.7	.3774	.4579	.4179
.8	.2653	.3236	.2979
.9	.2002	.2767	.2520
1.0	.1368	.1901	.1556

c) Performance of M- and H-group phrases

Recall	Addition of all phrases	Addition of all phrases and del- tion of H-terms	For comparison Thesaurus added (all classes)
.1	.8130	.8936	.8778
.2	.7574	.8021	.7540
.3	.6580	.6988	.6443
.4	.5712	.6514	.5488
.5	.5175	.5944	.5229
.6	.4366	.5164	.4620
.7	.3716	.4527	.3969
.8	.2658	.3146	.3268
.9	.2052	.2723	.2704
1.0	.1408	.1893	.1916

d) Performance of All Phrases

Phrase Performance for MED 450 Collection
(average precision at ten recall levels)

Table 9 (cont.)

Recall	Standard Word Stem (single terms)	Addition of H-group phrases	Addition of H-group phrases plus deletion of H-terms	Addition of H-group phrases, deletion of H-phrase components
.1	.7496	.7929	.8764	.8785
.2	.7071	.7535	.7882	.8076
.3	.6710	.6977	.7437	.7544
.4	.6452	.6483	.7063	.7157
.5	.6351	.6378	.6672	.6799
.6	.5866	.5870	.6592	.6593
.7	.5413	.5343	.5811	.5826
.8	.5004	.4994	.5521	.5454
.9	.3865	.3901	.4234	.4097
1.0	.3721	.3715	.4099	.3961

a) Performance with H-group phrases

Recall	Addition of M-group phrases	Addition of M-group phrases plus deletion of H-terms	Addition of M-group phrases, deletion of M-phrase components
.1	.7412	.7820	.7156
.2	.7339	.6847	.6912
.3	.6901	.6352	.6697
.4	.6698	.5961	.6300
.5	.6485	.5827	.5961
.6	.6075	.5019	.5535
.7	.5653	.4933	.5226
.8	.5425	.4462	.4794
.9	.4244	.3704	.3707
1.0	.4077	.3432	.3616

b) Performance with M-group phrases

Phrase Performance for TIME 425 Collection
(average precision at ten recall levels)

Table 10

Recall	Addition of M- and H-group phrases	Addition of M- and H-group phrases and deletion of H-terms	Addition of M- and H-group phrases and deletion of H-phrase components
.1	.7993	.8269	.8290
.2	.7922	.7824	.7859
.3	.7388	.7499	.7559
.4	.6877	.7288	.7347
.5	.6617	.6794	.6845
.6	.6187	.6624	.6615
.7	.5802	.6068	.6107
.8	.5448	.5674	.5623
.9	.4259	.4661	.4456
1.0	.4089	.4536	.4349

c) Performance of M- and H-group phrases

Recall	Addition of all phrases	Addition of all phrases, and deletion of H- terms	For comparison Thesaurus added (all classes)
.1	.7889	.8165	.7582
.2	.7819	.7893	.7305
.3	.7457	.7569	.6787
.4	.6877	.7298	.6478
.5	.6665	.6834	.6405
.6	.6208	.6624	.6237
.7	.5807	.6068	.5666
.8	.5448	.5674	.5213
.9	.4263	.4651	.4223
1.0	.4116	.4557	.4120

d) Performance of All Phrases

Phrase Performance for TIME 425 Collection
(average precision at ten recall levels)

Table 10 (cont.)

	CRAN 424	MED 450	TIME 425
Low Recall	Add M- and H-phrases; delete H-phrase components	Add M- and H-phrases, or add All; delete H-terms	Add H-phrases; delete H-phrase components
Medium Recall	Add All phrases; delete H-terms	Add M- and H-phrases; delete H-terms	Add M- and H-phrases; delete H-phrase components
High Recall	Add All phrases (no deletion)	Add M- and H-phrases; delete H-terms	Add M- and H-phrases, or add All; delete H-terms

Optimum Phrase Performance for Three Collections
at Three Recall Levels

Table 11

	CRAN 424	MED 450	TIME 425
Number of classes included in thesaurus	30	68	32
Number of thesaurus classes producing loss in discrimination value rank	3	20	9
Average rank in discrimination value order			
of a thesaurus class	365.3	1694.9	2093.0
of terms included in thesaurus class	1037.5	1432.6	2702.6
discrimination rank difference (DRD)	-672.2	+262.3	-609.6
Number of (nondistinct) thesaurus class identifiers added to query vectors			
a) ALL (all thesaurus classes)	41	76	42
b) NEG (classes with negative DRD)	35	50	29

Thesaurus Class Characteristics

Table 12

	CRAN 424	MED 450	TIME 425
Typical thesaurus categories producing improvement in discrimination value rank	<u>10017</u> viscid (1)L viscoelastic (1)M viscosity (22)M viscous (40)M viscous-interaction (1)L viscous-layer (1)L	<u>13015</u> septal (12)M septa (2)L septum (7)L transseptal (8)M	<u>8019</u> ideological (15)M ideologist (3)L ideology (5)L philosophy (6)L
Typical thesaurus categories producing loss in discrimination value rank	<u>10002</u> mass (17)M weight (10)M ----- <u>10020</u> shear (32)M	<u>13010</u> cerebrospinal (4)L cerebellar (2)L cerebral (22)H cranio-cerebral (1)L CSF (8)M	<u>8023</u> Asia (32)M Asian (15)M ----- <u>8024</u> democrat (41)M democracy (23)M

Typical Thesaurus Classes

(including document frequency for each term;

H, M, L high, medium, low frequency)

Table 13

R	Standard Word Stem (term frequency weighting)	Thesaurus Classes Added (ALL)	Thesaurus Classes Added (NEG)
.1	.6844	.7269	.7610
.2	.5303	.5833	.5840
.3	.4689	.5192	.4942
.4	.3482	.4219	.4205
.5	.3134	.3613	.3500
.6	.2556	.2867	.3144
.7	.1989	.2296	.2576
.8	.1631	.1756	.1992
.9	.1265	.1289	.1434
1.0	.1176	.1131	.1276

a) CRAN 424

R	Standard Word Stem (term frequency weighting)	Thesaurus Classes Added (ALL)	Thesaurus Classes Added (NEG)
.1	.7891	.8778	.8532
.2	.6750	.7540	.7471
.3	.5481	.6443	.6620
.4	.4807	.5488	.5285
.5	.4384	.5229	.5203
.6	.3721	.4620	.4421
.7	.3359	.3969	.3852
.8	.2195	.3268	.3098
.9	.1768	.2704	.2431
1.0	.1230	.1916	.1974

b) MED 450

Retrieval Performance for Thesaurus Runs
(average precision at ten recall points)

Table 14'

R	Standard Word Stem (term frequency weighting)	Thesaurus Classes Added (ALL)	Thesaurus Classes Added (NEG)
.1	.7496	.7582	.7474
.2	.7071	.7305	.7086
.3	.6710	.6787	.6721
.4	.6452	.6478	.6395
.5	.6351	.6405	.6323
.6	.5866	.6237	.6089
.7	.5413	.5666	.5550
.8	.5004	.5213	.5065
.9	.3865	.4223	.4018
1.0	.3721	.4120	.3911

c) TIME 425

Retrieval Performance for Thesaurus Runs
(average precision at ten recall points)

Table 14 (cont.)

ALL: all thesaurus classes used

NEG: only classes with negative DRD used

R	Standard Word Stem (term frequency weighting)	Thesaurus (all classes; no term deletion)	Phrases (all phrases; H component deletion (45))	Single Terms, Phrases and Thesaurus
.1	.6844	.7269	.8793	.8299
.2	.5303	.5833	.7344	.7652
.3	.4689	.5192	.6013	.6382
.4	.3482	.4219	.5205	.5341
.5	.3134	.3613	.4150	.4455
.6	.2556	.2867	.3623	.3708
.7	.1989	.2296	.3017	.2987
.8	.1631	.1756	.1953	.1974
.9	.1265	.1289	.1463	① { .1458
1.0	.1176	.1131	.1314	.1299

① All phrases, no term deletion is best

a) CRAN 424

R	Standard Word Stem (term frequency weighting)	Thesaurus (all classes; no term deletion)	Phrases (all phrases; H component deletion (18))	Single Terms, Phrases and Thesaurus
.1	.7891	.8778	.8603	.9002
.2	.6750	.7540	.9149	.8088
.3	.5481	.6443	.6903	② { .7092
.4	.4807	.5485	.6504	.6455
.5	.4384	.5229	.5937	.6272
.6	.3721	.4629	.5442	.5644
.7	.3359	.3969	.4719	.5294
.8	.2195	.3268	.3314	.4156
.9	.1768	.2704	.2807	.3594
1.0	.1230	.1916	.1991	.2548

② M- and H-group phrase addition, plus H-term deletion is best

b) MED 450

Retrieval Performance for Combined Phrase and Thesaurus Performance
(average precision at ten recall levels)
Table 15

R	Standard Word Stem (term frequency weighting)	Thesaurus (all classes, no term deletion)	Phrases (all phrases; H component deletion (49))	Single Terms, Phrases and Thesaurus
.1	.7496	.7582	.8488	③ (.9398
.2	.7071	.7305	.8419	.8138
.3	.6710	.6787	.7998	.7812
.4	.6452	.6478	.7729	.7681
.5	.6351	.6405	.7025	.7006
.6	.5866	.6237	.6800	.6882
.7	.5413	.5666	.6331	.6389
.8	.5004	.5213	.5895	.5915
.9	.3865	.4223	.4618	.4821
1.0	.3721	.4120	.4529	.4769

③ H-group phrase addition, plus H-term deletion is best

c) TIME 425

Retrieval Performance for Combined Phrase and Thesaurus Performance
(average precision at ten recall levels)

Table 15 (cont.)