A Theory of Term Importance in Automatic Text Analysis.

Most existing automatic content analysis and indexing techniques are based on work frequency characteristics applied largely in an ad hoc manner. Contradictory requirements arise in this connection, in that terms exhibiting high occurrence frequencies in individual documents are often useful for high recall performance (to retrieve many relevant items), whereas terms with low frequency in the whole collection are useful for high precision (to reject nonrelevant items). A new technique known as discrimination value analysis ranks the text words in accordance with how well they are able to discriminate the documents of a collection from each other; that is, the value of a term depends on how much the average separation between individual documents changes when the given term is assigned for content identification. The best words are those which achieve the greatest separation. The discrimination value analysis accounts for a number of important phenomena in the content analysis of natural language texts: (a) the role and importance of single words; (b) the role of juxtaposed words (phrases); (c) the role of word groups or classes, as specified in a thesaurus. Effective criteria can be given for assigning each term to one of these three classes, and for constructing optimal indexing vocabularies. (Author)
A Theory of Term Importance in Automatic Text Analysis
G. Salton*, C.S. Yang*, and C.T. Yu*

Abstract

Most existing automatic content analysis and indexing techniques are based on word frequency characteristics applied largely in an ad hoc manner. Contradictory requirements arise in this connection, in that terms exhibiting high occurrence frequencies in individual documents are often useful for high recall performance (to retrieve many relevant items), whereas terms with low frequency in the whole collection are useful for high precision (to reject nonrelevant items).

A new technique, known as discrimination value analysis ranks the text words in accordance with how well they are able to discriminate the documents of a collection from each other; that is, the value of a term depends on how much the average separation between individual documents changes when the given term is assigned for content identification. The best words are those which achieve the greatest separation.

The discrimination value analysis accounts for a number of important phenomena in the content analysis of natural language texts:

a) the role and importance of single words;
b) the role of juxtaposed words (phrases);
c) the role of word groups or classes, as specified in a thesaurus.

Effective criteria can be given for assigning each term to one of these three classes, and for constructing optimal indexing vocabularies.

* Department of Computer Science, Cornell University, Ithaca, N.Y. 14850.
+ Department of Computer Science, University of Alberta, Edmonton, Alberta.

1. Document Space Configuration

Consider a collection of entities \( D \) (documents) represented by weighted properties \( w \). In particular, let

\[
D_i = (w_{i1}, w_{i2}, ..., w_{it})
\]

where \( w_{ij} \) represents the weight of term \( j \) in the vector corresponding to the \( ith \) document. Given two documents \( D_1 \) and \( D_2 \), it is possible to define a measure of relatedness \( s(D_1, D_2) \) between the documents depending on the similarity of their respective term vectors. In three dimensions (when only three terms identify the documents), the situation may be represented by the configuration of Fig. 1, where the similarity between any two of the document vectors may be assumed to be a function inversely related to the angle between them. That is, when two document vectors are exactly the same, the corresponding vectors are superimposed and the angle between them is zero.

When the dimensionality of the space exceeds three, that is when more than three terms are used to identify a given document, the envelope of the vector space may be used to represent the collection as in the example of Fig. 2. Here only the tips of the document vectors are shown, represented by \( x \)'s, and the distance between two \( x \)'s is inversely related to the similarity between the corresponding document vectors — the smaller the distance between \( x \)'s, the smaller will be the angle between the vectors, and thus the more similar the term assignment.

The theory is validated by citing experimental results.
A central document, or centroid $C$, may be introduced, located in the center of the document space, which for certain purposes may represent the whole collection. The $i$th vector element $c_i$ of the centroid can simply be defined as the average of the $i$th term $w_{ij}$ across the $n$ documents of the collection; that is

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

It is clear that a particular document space configuration, such as that of Fig. 2, reflects directly the details of the indexing chosen for the identification of the documents. This raises the question about the choice of an optimum indexing process, or alternatively, about an effective document space configuration. A number of studies, carried out over the last few years, indicate that a good document space is one which maximizes the average separation between pairs of documents. In particular, the document space will be maximally separated, when the average distance between each document and the space centroid is maximized, that is, when

$$Q = \sum_{i=1}^{n} s(C, b_i)$$ 

is minimum. Obviously, in such a case, it may be easy to retrieve each given document without also necessarily retrieving its neighbors. This insures a high precision output, since the retrieval of a given relevant item will then not also entail the retrieval of many nonrelevant items in its vicinity.

Furthermore, when the relevant documents are located in the same general area of the space, high recall may also be obtainable, since many relevant items may then be correctly retrieved, and many nonrelevant correctly rejected.

A particular indexing system, known as the discrimination value model, assigns the highest weight, or value, to those terms which cause the maximum possible separation between the documents of a collection. This model is described and analyzed in the remainder of this study.

2. The Discrimination Value Model.

The discrimination value of a term is a measure of the changes in space separation which occur when a given term is assigned to a collection of documents. A good discriminator is one which when assigned as an index term will render the documents less similar to each other; that is, its assignment decreases the space density. Contrariwise, a poor discriminator increases the density of the space. By computing the space densities both before and after assignment of each term, it is possible to rank the terms in decreasing order of their discrimination values.

In particular, consider a measure of the space density, such as the $Q$ value given in equation (2), and let $Q_k$ represent the density $Q$ with the $k$th term removed from all document (and from the centroid) vectors. The discrimination

 Retrieval performance is often measured by parameters such as recall and precision, reflecting the ratio of relevant items actually retrieved, and of retrieved items actually relevant.
value of term \( k \) may then be defined as

\[ DV_k = Q_k - Q. \]  

(3)

Obviously, if term \( Q \) is a good discriminator, then its removal will cause a compression in the document space (an increase in space density), because its assignment would have resulted in an increase in space separation. Thus for good discriminators, \( Q_k \geq Q \) and \( DV_k \) is positive. The reverse is true for poor discriminators whose removal causes a decrease in space density, leading to negative discrimination values. A vast majority of the terms may be expected to produce neither increase nor decrease in space density; in such a case a discrimination value near zero is obtained. The operations of a good discriminator are illustrated in the simplified drawing of Fig. 3.

In the retrieval experiments conducted earlier with three collections in aeronautics (Tranfield collection, 426 documents comprising 2651 distinct terms), medicine (Medline collection, 450 documents comprising 4726 terms), and world affairs (Time collection, 425 documents comprising 14098 terms), the discrimination value model produced excellent retrieval results. (1) In particular, a term weighting system which assigns to each term \( k \) a value \( w_{kj} \) consisting of the product of its frequency of occurrence in document \( j \), \( f_{kj} \), multiplied by its discrimination value \( DV_k \),

\[ w_{kj} = f_{kj} \cdot DV_k. \]  

(4)

produces recall and precision improvements of about ten percent over methods where only the term frequencies \( f_{kj} \) are taken into account.*

It may be of interest to inquire what kind of terms are favored by a weighting system such as that of expression (4), and what accounts for the value of the discrimination model. Some experimental evidence relating the discrimination values to certain frequency characteristics of the terms in the document collections is presented in the next section. This in turn, leads to an indexing theory to be examined in the remainder of this study.

3. Discrimination Values and Document Frequencies

Consider any term \( k \) assigned to a collection of documents, and let \( c_k \) be its document frequency, defined as the number of documents in the collection to which term \( k \) is assigned. More specifically,

\[ d_k = \sum_{j=1}^{n} b_{kj}, \]

where \( b_{kj} = 1 \) whenever \( f_{kj} \geq 1 \), and \( b_{kj} = 0 \) otherwise. It is instructive to arrange the terms assigned to a document collection into disjoint sets in such a way that the terms assigned to a given set have equal document frequencies \( d_k = 1 \). Moreover, for each such set of terms the average rank in decreasing discrimination value order may be computed, thereby relating document frequencies with discrimination values.†

---

* Terms receiving high weights according to expression (4) are those which exhibit high occurrence frequencies in certain specified documents, and at the same time can distinguish these documents from the remainder of the collection.

† For a set of \( t \) terms, the discrimination value rank ranges from 1 for the best discriminator to \( t \) for the worst.
A plot giving the average discrimination value rank for the terms exhibiting certain document frequency ranges is shown in Figs. 4(a), (b), and (c) for the collections in aerodynamics, medicine, and world affairs (Cranfield, Medlars, and Time) respectively. It may be seen that a U-shaped curve is obtained in each case, with the following interpretation:

a) the terms with very low document frequencies, located on the left-hand side of Fig. 4 are poor discriminators, with average discrimination value ranks in excess of $t/2$ for $t$ terms;

b) the terms with high document frequencies exceeding $n/10$, located on the right-hand side of Fig. 4 are the worst discriminators, with average discrimination value 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 value ranks are generally below $t/5$.

The output of Fig. 4 shows average discrimination value ranks only. Before deciding that all terms with low and high document frequencies can automatically be disregarded, it is useful to determine whether any good discriminators are in fact included in the corresponding low frequency and high frequency term sets.

Figs. 5(a) and 5(b) show sets of low frequency terms for the Medlars and Time collections respectively, together with the number of good discriminators — those with discrimination rank between 1 and 100 — included in each set. Fig. 5 shows overlapping term sets, consisting of all terms with document frequency equal to 1, 1 and 2, 1 to 3, etc., together with the percentage figures of the total number of terms represented by the corresponding sets.

Thus when seventy percent of the terms are taken in increasing document frequency order — corresponding in the Medlars collection to about 3700 terms out of 4700 with document frequencies of 1 or 2, and in the Time collection to 9000 terms out of 14000 with document frequencies 1 to 3 — it is seen that only about 15 good discriminators are included for Medlars, and about 12 for Time. When the proportion of terms increases to eighty percent in increasing document frequency order, including 3000 Medlars terms, or 11000 Time terms, ranging in document frequency from 1 to 5, the number of good discriminators rises to 30 for Medlars and to 35 for Time. When so few good terms are included among the mass of low frequency terms, it is obvious that special provisions must be made in any indexing process for the utilization of these terms.

Consider now the very high frequency terms — those which according to the output of Fig. 4 exhibit the lowest discrimination values. While the number of such terms is not large, each of the terms accounts for a substantial portion of the total term assignments to the documents of a collection because of the document frequency involved.

The output of Fig. 6(a) for Medlars, and 6(b) for Time shows that about four percent of the high-frequency terms present in a document collection, accounts for forty to fifty percent of all term assignments, when the terms are taken in decreasing document frequency order. The absolute number of distinct terms is 200 approximately for the Medlars collection and about 500 for Time. In each case, less than 15 of these terms are classified as good discriminators. When the proportion of terms taken in high frequency order increases to six percent, accounting for 56 percent of the term assignments in Medlars, and 57 percent for Time, the number of good discriminators increases to about 20 in each case.
The information included in Figs. 5 and 6 is summarized in Table 1. In each case, certain cutoff percentages are given for terms taken either in low document frequency or in high document frequency order. For each such percentage, the number of good discriminators included in the corresponding term set is stated for each of the three test collections. Thus when sixty percent of the terms are taken in increasing document frequency order, not a single good discriminator is included among the 1668 terms for the Cranfield collection; only 5 of the top 50 terms, or 16 of the top 100, are present among the 3238 Medlars terms; finally, for Time, 11 out of the top 50, or 11 of the top 100 are included among the first 891 low frequency terms.

The number of good discriminators included among the high frequency terms for the three collections is similarly low, as shown in the bottom half of Table 1.

The conclusion to be reached from the data of Figs. 5 and 6 and of Table 1 is that very few good discriminators are included among the bottom seventy percent, or among the top four percent when the terms included in a collection of documents are taken in increasing document frequency order. This fact is used to construct an indexing strategy in the remainder of this study.

4. A Strategy for Automatic Indexing

Consider the graph of Fig. 7 in which the terms are once again arranged in increasing document frequency order. If the assumption is correct that the best terms for indexing purposes are concentrated in the set whose document frequency is neither too high nor too low — the frequency being approximately between \( \frac{n}{100} \) and \( \frac{n}{10} \) — then the following term transformations should be undertaken:

a) Terms whose document frequency lies between \( \frac{n}{100} \) and \( \frac{n}{10} \) should be used for indexing purposes directly without any transformation; these terms include the vast majority of the good discriminators.

b) Terms whose document frequency is too high — above \( \frac{n}{10} \) — 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 items that are extraneous). These terms should be transformed into lower frequency terms — right-to-left on the graph of Fig. 7 — thereby enhancing the precision performance.

c) Terms whose document frequency is too low — below \( \frac{n}{100} \) — are so rare and specific that they cannot retrieve an acceptable proportion of the documents relevant to a given query; hence their use depresses the recall performance. These terms should be transformed into higher frequency terms — left-to-right on the graph of Fig. 7 — thereby enhancing the recall performance.

It remains to describe the right-to-left and left-to-right transformations that may be used to generate useful indexing vocabularies. The obvious way of transforming the high frequency terms into lower frequency entities is to combine them into indexing phrases. In general, a phrase such as "programming language" exhibits a lower assignment frequency than either of the high frequency components "language" or "program". The summary of Fig. 7 then indicates that:
Indexing phrases should be constructed from high frequency single term components in order to enhance the precision performance of the retrieval system.

The other left-to-right transformation which is required for recall enhancing purposes is now equally obvious. Low frequency terms with somewhat similar properties, or meanings, can be combined into term classes, normally specified by a thesaurus of related terms, or synonym dictionary. When a single term is replaced for indexing purposes by a thesaurus class consisting of several terms, the assignment frequency of the thesaurus class will in general exceed that of any of the components included in the class. Thus:

The main virtue of a thesaurus is its ability to group a number of low frequency terms into thesaurus classes, thereby enhancing the recall performance.

A large number of different strategies is available for the generation of indexing phrases and term thesauruses. Consider first the criteria used for the formation of phrases. A phrase might be created whenever two or more components cooccur in the same document, or query; or when they cooccur in the same paragraph, or sentence of a document; or when they occur in certain specified positions within the same sentences; or, finally, when they cooccur in certain specified positions in a text while exhibiting certain predetermined syntactical relationships. The methods needed to identify the indexing phrases attached to a given document or query may then range from quite simple (any pair of noncommon terms cooccurring in a document may represent a phrase) to quite complex (the various phrase components must exhibit appropriate syntactical relationships, and these relationships must be ascertained). [3]

For present purposes, a compromise position is adopted which bypasses an expensive syntactic analysis system in favor of the following procedure:

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 occurs between components), and provided that at least one of the components of each phrase is a high-frequency term;

d) phrases for which both components are identical are eliminated;

e) duplicate phrases, where all components match an already existing phrase are eliminated.

The texts of all documents are checked for the presence of any phrase thus defined from the query statements, and appropriate weights are assigned.

The phrase formation process is illustrated in Fig. 8 for a query dealing with world affairs. It is seen that this query gives rise to eight distinct phrases with adjacent components, plus seven additional phrases for which the components are separated by one intervening word in the reduced query text.

It remains to determine an appropriate weight to be assigned to each phrase created by the foregoing process. Thus if terms $p$ and $q$ exhibit weights $w_{pq}$ and $w_{q}$, respectively in document $i$, corresponding, for example to the frequencies of occurrence of the respective terms in the document, the phrase consisting of components $p$ and $q$ might be assigned weight $w_{pq}$ defined as
A somewhat more refined weighting method uses \( w_{pq} \) in conjunction with an "inverse document frequency" (IDF) factor which gives higher weights to phrases that occur comparatively rarely in the collection. The original inverse document frequency (IDF) factor, introduced by Sparck Jones, was defined as:[1]

\[
\text{IDF}_k = \log \frac{n}{d_k} \cdot \frac{1}{\log \frac{q}{r}}
\]

where \( \text{IDF}_k \) is the IDF factor for term \( k \), and \( d_k \) is the document frequency of term \( k \) in a collection of \( n \) documents. Clearly \( \text{IDF}_k \) is large when \( d_k \) is small, and becomes small as \( d_k \) approaches \( n \).

By analogy, a phrase IDF factor may be defined as:

\[
\text{IDF}_{pq} = \log \frac{n}{d_p + d_q}.
\]

where \( d_p \) and \( d_q \) are the respective document frequencies of phrase components \( p \) and \( q \).

In conformity with the composite weighting system of equation (4) which uses the product of term frequencies and discrimination values, a composite phrase weight \( w_{ipq} \) for phrase \( pq \) in document \( i \) may then be defined as the product of the IDF factor and the average component weight (equations (5) and (6)).

\[
w_{ipq} = \frac{w_p + w_q}{2}.
\]

In a retrieval environment, the phrases defined by the foregoing procedure may be used to replace the original phrase components — that is, the original components may be removed from the document and query vectors before the phrase identifiers are added. Alternatively, phrase components may be used in addition to the single term components. For the experiments described in the next section, the former policy was used in that phrases are introduced replacing original component terms.

Consider now the converse to the right-to-left phrase formation process, namely the left-to-right thesaurus construction method. Here the notion is to use low frequency terms and to assemble them into classes of terms replacing the original vector components. If \( d_p \) and \( d_q \) are the document frequencies of terms \( p \) and \( q \) respectively, the document frequency of the class which includes both \( p \) and \( q \) may be defined as

\[
\text{D}_{pq} = d_p + d_q - d_{pq}.
\]

In general \( D_{pq} \) may be expected to be larger than either \( d_p \) or \( d_q \) individually. When \( n \) terms are included in a given term class, the document frequency of the class is defined simply as the number of documents in which at least one term associated to that class appears.

As before, the weighting system of expression (7) assigns high weights to phrases with highly weighted components in individual documents but with...
Term classes are often defined by a thesaurus, and a given thesaurus class normally includes terms that are sufficiently similar in meaning, or context, to make it reasonable to ignore their differences for indexing purposes. A great many thesaurus construction procedures have been described in the literature including manual term grouping as well as fully automatic methods. Among the latter are the so-called associative indexing procedures, where statistically associated terms are jointly assigned to the documents of a collection, and a variety of term clustering methods designed to group into a common class those terms which exhibit similar term assignments to the documents of a collection.

For experimental purposes it may be sufficient to use existing manually constructed thesauruses for the three test collections, and restricting the thesaurus to include only classes whose document frequency does not exceed a stated maximum. Such a thesaurus then effectively limits the number of high-frequency terms than can appear in any class, and provides the left-to-right frequency transformation specified by the model of Fig. 7. The weight with which a thesaurus class is assigned to a document or query vector may be defined as the average weight of the component terms originally present in that vector.

A frequency-restricted thesaurus such as the one described above may not specify classes that are completely identical with the term classes obtainable by initially using only the low frequency terms for a separate term clustering process; however the experimental recall-precision results may be expected to be close to those produced by an original thesaurus construction method.

The recall-precision results obtained from the operations modelled in Fig. 7 are examined in the next section.

5. Experimental Results
The right-to-left phrase formation process is designed to produce lower frequency entities from high frequency components, and vice versa for the left-to-right thesaurus grouping process. The data of Table 2 prove that the required frequency alterations are in fact obtained by the two transformations for the test collections in use.

Table 2(a) shows that the document frequency of the phrases is only about one third as large as the frequency of the individual components entering the phrase formation process. In Table 2(b) the reverse is seen to be the case for the thesaurus concepts whose document frequency is one and a half times that of the individual thesaurus entries. If the model of Fig. 7 specifying ideal frequency characteristics for index terms is appropriate, considerably better recall and precision output should be obtainable with the transformed terms (phrases and thesaurus classes) than the originals.

Detailed recall-precision output is contained in Tables 3 and 4, and in the summary in Table 5 for the various indexing methods applied to the three test collections in aerodynamics, medicine, and world affairs. Performance figures comparing the standard term frequency weighting \(f(k,i)\) for single terms \(k\) in documents \(i\) with the phrase process are shown in Table 3. The phrase procedure uses the normal single terms in addition to indexing phrases weighted in accordance with the formula of expression (7).

Table 3 contains precision figures averaged over 10 user queries for each of the test collections at ten specified recall levels ranging in magnitude from 0.1 to 1.0 in steps of 0.1. The percentage improvement in precision for the phrase
process over the standard is also given at each recall level, together with an average improvement ranging from a high of 39 percent for the Medlars collection to a low of 17 percent for Time.

Table 4 contains output similar to that already shown in Table 3. However, the data in Table 4 apply to an indexing system using both left-to-right (thesaurus) and right-to-left (phrase) transformations. It is seen from Table 4 that the thesaurus transformation adds an additional average improvement of 13 percent in precision for the Medlars collection; additional advantages are also obtained for the Cranfield and Time collections.

The evaluation results are summarized in Table 5. It is seen that average precision values of approximately 0.70, 0.60, and 0.50 at high, medium, and low precision are transformed into average figures of 0.90, 0.80, and 0.70 approximately when the discrimination properties of the terms are optimized. The retrieval results displayed in Tables 3, 4, and 5 have not been surpassed by any manual or automatic indexing procedures previously tried with sample document collections and user queries. Furthermore, because of the high average precision values produced by the indexing theories described in this study, it is not likely that additional drastic improvements in retrieval effectiveness are obtainable in the foreseeable future.

References

The vector representation of the document space is shown in Fig. 1. The centroid of the space is denoted by the origin, and the individual documents are represented by points in the multidimensional space. In Fig. 2, the centroid of the space is marked by a circle, and the individual documents are shown as smaller circles.
Before Assignment of Term

After Assignment of Term

Operation of Good Discriminating Term

Fig. 3

Cranfield Collection
424 Documents
2651 Terms

Average Discrimination Value Rank of Terms

Document
Main Centroid
Medlars Collection
450 Documents
4726 Terms

Time 425 Collection
14,098 terms*

Average Discrimination Value Rank of Terms

Fig. 4(b)
Number of Top 100 Discriminators in Range

Number of Good Discriminators Among Low Frequency Terms

Fig. 5(a)

Number of Top 100 Discriminators in Range

Number of Good Discriminators Among Low Frequency Terms

Fig. 5(b)
QUERY: COALITION GOVERNMENT TO BE FORMED IN ITALY BY THE LEFT WING SOCIALISTS, THE REPUBLICANS, SOCIALISTS, AND CHRISTIAN DEMOCRATS.

DELETING COMMON WORDS AND ELIMINATING SUFFIXES:

COALITION GOVERN, GOVERN FORM, ITALY, ITALY LEFT, LEFT.

PHRASES:

ADJACENT COMPONENTS:

COALITION GOVERN, GOVERN FORM, ITALY, ITALY LEFT.

Duplicate Phrases:

Fig. 7

Summarization of Discrimination Values of Terms in Frequency Range

First, Second, Most Discriminates

(60% of Terms)

Frequency (in Documents)
<table>
<thead>
<tr>
<th>Type of Terms</th>
<th>Fraction of Terms Covered (Fraction of Term Assignments)</th>
<th>Number of Terms</th>
<th>Document Frequency of Terms</th>
<th>Number of Good Discriminators</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>CRAN 1669</td>
<td>1-3</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>MED 2320</td>
<td>1-2</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td></td>
<td>TIME 8916</td>
<td>1-2</td>
<td>1</td>
</tr>
<tr>
<td>Low Frequency</td>
<td></td>
<td>CRAN 1999</td>
<td>1-4</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>MED 3602</td>
<td>1-3</td>
<td>8</td>
</tr>
<tr>
<td></td>
<td></td>
<td>TIME 9964</td>
<td>1-3</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td></td>
<td>CRAN 2153</td>
<td>1-9</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td></td>
<td>MED 3085</td>
<td>1-4</td>
<td>8</td>
</tr>
<tr>
<td></td>
<td></td>
<td>TIME 11304</td>
<td>1-6</td>
<td>12</td>
</tr>
<tr>
<td>High Frequency</td>
<td></td>
<td>CRAN 91</td>
<td>5-214</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td></td>
<td>MED 163</td>
<td>20-130</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td></td>
<td>TIME 402</td>
<td>36-271</td>
<td>8</td>
</tr>
<tr>
<td></td>
<td></td>
<td>CRAN 105</td>
<td>48-214</td>
<td>7</td>
</tr>
<tr>
<td></td>
<td></td>
<td>MED 192</td>
<td>18-130</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td></td>
<td>TIME 555</td>
<td>33-271</td>
<td>8</td>
</tr>
<tr>
<td></td>
<td></td>
<td>CRAN 115</td>
<td>44-214</td>
<td>9</td>
</tr>
<tr>
<td></td>
<td></td>
<td>MED 206</td>
<td>17-130</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td></td>
<td>TIME 680</td>
<td>29-271</td>
<td>8</td>
</tr>
</tbody>
</table>

Number of Good Discriminators for Various Deletion Percentage of Low and High Frequency Terms

Table 1

<table>
<thead>
<tr>
<th>Minimum Document Frequency needed for High-Frequency Component</th>
<th>Average Document Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>Single Terms Entering Phrase Process</td>
<td>Phrases</td>
</tr>
<tr>
<td>CRANFIELD (45)</td>
<td>126</td>
</tr>
<tr>
<td>MEDLARS (22)</td>
<td>49</td>
</tr>
<tr>
<td>TIME (49)</td>
<td>101</td>
</tr>
</tbody>
</table>

Average Document Frequency for Phrases
Table 2(a)

<table>
<thead>
<tr>
<th>Maximum Document Frequency needed for Thesaurus Class to Inclusion</th>
<th>Average Document Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>Single Terms Entering Thesaurus Process</td>
<td>Thesaurus Classes</td>
</tr>
<tr>
<td>CRANFIELD (60)</td>
<td>26</td>
</tr>
<tr>
<td>MEDLARS (40)</td>
<td>10</td>
</tr>
<tr>
<td>TIME (60)</td>
<td>17</td>
</tr>
</tbody>
</table>

Average Document Frequency for Thesaurus Classes
Table 2(b)
<table>
<thead>
<tr>
<th>Standard Term Frequency</th>
<th>Phrase Assignment</th>
<th>Average</th>
<th>Standard Term Frequency</th>
<th>Phrase Assignment</th>
<th>Average</th>
<th>Standard Term Frequency</th>
<th>Phrase Assignment</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

**Table 3**

Average Precision Values at Ten Recall Points

(Phrase Process vs. Standard)

---

<table>
<thead>
<tr>
<th>Standard Term Frequency</th>
<th>Thesaurus Plus Phrases</th>
<th>Average</th>
<th>Standard Term Frequency</th>
<th>Thesaurus Plus Phrases</th>
<th>Average</th>
<th>Standard Term Frequency</th>
<th>Thesaurus Plus Phrases</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

**Table 4**

Average Precision Values at Ten Recall Points

(Phrase and Phrases vs. Standard)
<table>
<thead>
<tr>
<th>Collection</th>
<th>Frequency Comparison</th>
<th>Plus Thesaurus Comparison</th>
</tr>
</thead>
<tbody>
<tr>
<td>TIME 425</td>
<td>Automatic Phrases vs. Standard Term +17%</td>
<td>Automatic Phrases vs. Standard Run +16%</td>
</tr>
<tr>
<td>MED 460</td>
<td>Automatic Phrases vs. Standard Term +39%</td>
<td>Automatic Phrases vs. Standard Run +52%</td>
</tr>
<tr>
<td>CRAN 424</td>
<td>Automatic Phrases vs. Standard Term +32%</td>
<td>Automatic Phrases vs. Standard Run +37%</td>
</tr>
</tbody>
</table>

Summary of Recall-Precision Evaluation (Three Collections)

Table 5