74-218

A Vector Space Model for Automatic Indexing

G. Salton, A. Wong, and C.S. Yang [+]

Abstract

In a document retrieval, or other pattern matching environment where stored entities (documents) are compared with each other, or with incoming patterns (search requests), it appears that the best indexing (property) space is one where each entity lies as far away from the others as possible; that is, retrieval performance correlates inversely with space density. This result is used to choose an optimum indexing vocabulary for a collection of documents. Typical evaluation results are shown demonstrating the usefulness of the model.

1. Document Space Configurations

Consider a document space, consisting of documents D_i, each identified by one or more index terms T_j; the terms may be weighted according to their importance, or unweighted with weights restricted to 0 and 1.* A typical three-dimensional index space is shown in Fig. 1, where each item is identified by up to three distinct terms. The three-dimensional example may be extended to t dimensions when t different index terms are present. In that case, each document D_i is represented by a t-dimensional vector

$$D_i = (d_{i1}, d_{i2}, \ldots, d_{it}),$$

d_{ij} representing the weight of the jth term.

[+]Department of Computer Science, Cornell University, Ithaca, N.Y., 14850

*Although we speak of documents and index terms, the present development applies to any set of entities identified by weighted property vectors.

Given the index vectors for two documents, it is possible to compute a similarity coefficient between them $s(D_i, D_j)$, reflecting the degree of similarity in the corresponding terms and term weights. Such a similarity measure might be an inverse function of the angle between the corresponding vector pairs — when the term assignment for two vectors is identical, the angle will be zero producing a maximum similarity measure.

Instead of representing each document by a complete vector originating at the 0-point in the coordinate system, the relative position of the vectors is preserved by considering only the envelope of the space. In that case, each document is graphically identified by a single point whose position is specified by the area where the corresponding document vector touches the envelope of the space. Two documents with similar index terms are then represented by points that are verv close together in the space: obviously the distance between two document points in the space is inversely correlated with the similarity between the corresponding vectors.

Since the configuration of the document space is a function of the manner in which terms and term weights are assigned to the various documents of a collection, one may ask whether an optimum document space configuration exists, that is, one which produces an optimum retrieval performance.*

*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. The question concerning optimum space configurations may then be more conventionally expressed in terms of the relationship between document indexing on the one hand, and retrieval performance on the other.

If nothing special is known about the documents under consideration, one might conjecture that an ideal document space is one where documents that are jointly relevant to certain user queries are clustered together, thus insuring that they would be retrievable jointly in response to the corresponding queries. Contrariwise, documents that are never wanted simultaneously would appear well separated in the document space. Such a situation is depicted in the illustration of Fig. 2, where the distance between two x's representing two documents is inversely related to the similarity between the corresponding index vectors.

While the document configuration of Fig. 2 may indeed represent the best possible situation, assuming that relevant and nonrelevant items with respect to the various queries are separable as shown, no practical way exists for actually producing such a space, because during the indexing process, it is difficult to anticipate what relevance assessments the user population will provide over the course of time. That is, the optimum configuration is difficult to generate in the absence of a priori knowledge of the complete retrieval history for the given collection.

In these circumstances, one might conjecture that the next best thing is to achieve a maximum possible separation between the individual documents in the space, as shown in the example of Fig. 3. Specifically, for a collection of n documents, one would want to minimize the function

$$F = \sum_{i=1}^{n} \sum_{\substack{j=1 \\ i \neq j}}^{n} s(D_i, D_j), \qquad (1)$$

where $s(D_i, D_j)$ is the similarity between documents i and j. Obviously when the function of equation (1) is minimized, the average similarity between document pairs is smallest, thus guaranteeing that each given document may be retrieved when located sufficiently close to a user query <u>without also necessarily retrieving its neighbors</u>. This insures a high precision search output, since a given relevant item is then retrievable without also retrieving a number of nonrelevant items in its vicinity. In cases where several different relevant items for a given query are located in the same general area of the space, it may then also be possible to retrieve many of the relevant while rejecting most of the nonrelevant. This produces both high recall and high precision.*

Two questions then arise: first, is it in fact the case that a separated document space leads to a good retrieval performance, and vice-versa that improved retrieval performance implies a wider separation of the documents in the space; second, is there a practical way of measuring the space separation. In practice, the expression of equation (1) is difficult to compute since the number of vector comparisons is proportional to n^2 for a collection of n documents.

---

*In practice, the best performance is achieved by obtaining for each user a desired recall level (a specified proportion of the relevant items); at that recall level, one then wants to maximize precision by retrieving as few of the nonrelevant as possible.

For this reason, a clustered document space is best considered, where the documents are grouped into classes, each class being represented by a class centroid. A typical clustered document space is shown in Fig. 4, where the various document groups are represented by circles and the centroids by black dots located more or less at the center of the respective clusters.* For a given document class K comprising m documents, each element of the centroid C may then be defined as the average weight of the same elements in the corresponding document vectors, that is

$$c_j = \frac{1}{m} \sum_{\substack{i=1 \\ D_i \epsilon K}}^{m} d_{ij}. \qquad (2)$$

Corresponding to the centroid of each individual document cluster, a centroid may be defined for the whole document space. This main centroid, represented by a small rectangle in the center of Fig. 4, may then be obtained from the individual cluster centroids in the same manner as the cluster centroids are computed from the individual documents. That is, the main centroid of the complete space is simply the average of the various cluster centroids.

---

*A number of well-known clustering methods exist for automatically generating a clustered collection from the term vectors representing the individual documents. [1]

In a clustered document space, the space density measure consisting of the sum of all pairwise document similarites, introduced earlier as equation (1), may be replaced by the sum of all similarity coefficients between each document and the main centroid, that is

$$Q = \sum_{i=1}^{n} s(C^*, D_i), \qquad (3)$$

where C* denotes the main centroid. Whereas the computation of equation (1) requires n^2 operations, an evaluation of equation (3) is proportional to n.

Given a clustered document space such as the one shown in Fig. 4, it is necessary to decide what type of clustering represents most closely the separated space shown for the unclustered case in Fig. 3. If one assumes that documents that are closely related within a single cluster normally exhibit identical relevance characteristics with respect to most user queries, then the best retrieval performance should be obtainable with a clustered space exhibiting tight individual clusters, but large intercluster distances; that is,

a) the average similarity between pairs of documents within a single cluster should be maximized, while simultaneously

b) the average similarity between different cluster centroids is minimized.

The reverse obtains for cluster organizations not conducive to good performance where the individual clusters should be loosely defined, whereas the distance between different cluster centroids should be small.

In the remainder of this study, actual performance figures are given relating document space density to retrieval performance, and conclusions are reached regarding good models for automatic indexing.

2. Correlation between Indexing Performance and Space Density

The main techniques useful for the evaluation of automatic indexing methods are now well understood. In general, a simple straightforward process can be used as a base-line criterion — for example, the use of certain word stems extracted from documents or document abstracts, weighted in accordance with the frequency of occurrence (f_i^k) of each term k in document i. This method is known as term-frequency weighting. Recall-precision graphs can be used to compare the performance of this standard process against the output produced by more refined indexing methods. Typically, a recall-precision graph is a plot giving precision figures, averaged over a number of user queries, at ten fixed recall levels, ranging from 0.1 to 1.0 in steps of 0.1. The better indexing method will of course produce higher precision figures at equivalent recall levels.

One of the best automatic term weighting procedures evaluated as part of a recent study consisted of multiplying the standard term frequency weight f_i^k by a factor inversely related to the document frequency d_k of the term (the number of documents in the collection to which the term is assigned). [2] Specifically, if d_k is the document frequency of term k, the inverse document frequency IDF_k of term k may be defined as [3]:

$$(\text{IDF})_k = \lceil \log_2 n \rceil - \lceil \log_2 d_k \rceil + 1.$$

A term weighting system proportional to $(f_i^k \cdot \text{IDF}_k)$ will assign the largest weight to those terms which arise with high frequency in individual documents, but are at the same time relatively rare in the collection as a whole.

It was found in the earlier study that the average improvement in recall and precision (average precision improvement at the ten fixed recall points) was about 14 percent for the system using inverse document frequencies over the standard term frequency weighting. The corresponding space density measurements are shown in Table 1 using two different cluster organizations for a collection of 424 documents in aerodynamics:

a) Cluster organization A is based on a large number of relatively small clusters. and a considerable amount of overlap between the clusters (each document appears in about two clusters on the average); the clusters are defined from the document-query relevance assessments, by placing into a common class all documents jointly declared relevant to a given user query.

b) Cluster organization B exhibits fewer classes (83 versus 155) of somewhat larger size (6.6 documents per class on the average versus 5.8 for cluster organization A); there is also much less overlap among the clusters (1.3 clusters per document versus 2.1). The classes are constructed by using a fast automatic tree-search algorithm due to Williamson. [4]

A number of space density measures are shown in Table 1 for the two cluster organizations, including the average similarity between the documents and the corresponding cluster centroids (factor x); the average similarity between the cluster centroids and the main centroid; and the average similarity between pairs of cluster centroids (factor y). Since a well-separated space corresponds to tight clusters (large x) and large differences between different clusters (small y), the ratio y/x can be used to measure the overall space density. [5]

It may be seen from Table 1, that all density measures are smaller for the indexing system based on inverse document frequencies; that is, the documents within individual clusters resemble each other less, and so do the complete clusters themselves. However, the "spreading out" of the clusters is greater than the spread of the documents inside each cluster. This accounts for the overall decrease in space density between the two indexing systems. The results of Table 1 would seem to support the notion that improved recall-precision performance is associated with decreased density in the document space.

The reverse proposition, that is, whether decreased performance implies increased space density may be tested by carrying out term weighting operations inverse to the ones previously used. Specifically, since a weighting system in <u>inverse</u> document frequency order produces a high recall-precision performance, a system which weights the terms directly in order of their document frequencies (terms occurring in a large number of documents receive the highest weights) should be correspondingly poor. In the output of Table 2, a term weighting system proportional to ($f_i^k \cdot DF_k$) is used, where f_i^k is again the term frequency of term k in document i, and DF_k is defined as $10/(IDF)_k$. The recall-precision figures of Table 2 show that such a weighting system produces a decreased performance of about ten percent, compared with the standard.

The space density measurements included in Table 2 are the same as those in Table 1. For the indexing system of Table 2, a general "bunching up" of the space is noticeable, both inside the clusters and between clusters. However, the similarity of the various cluster centroids increases more than that between documents inside the clusters. This accounts for the higher y/x factor by 16 and 7 percent for the two cluster organizations, respectively.

3. Correlation between Space Density and Indexing Performance

In the previous section it was shown that certain indexing methods which operate effectively in a retrieval environment are associated with a decreased density of the vectors in the document space, and contrariwise that poor retrieval performance corresponds to a space that is more compressed.

The relation between space configuration and retrieval performance may, however, also be considered from the opposite viewpoint. Instead of picking document analysis and indexing systems with known performance characteristics and testing their effect on the density of the document space, it is possible artifically to change the document space configurations in order to ascertain whether the expected changes in recall and precision are in fact produced.

The space density criteria previously given stated that a collection of small tightly clustered documents with wide separation between individual clusters should produce the best performance. The reverse is true of large nonhomogeneous clusters that are not well separated. To achieve improvements in performance, it would then seem to be sufficient to increase the similarity between document vectors located in the same cluster, while decreasing the similarity between different clusters or cluster centroids. The first effect is achieved by emphasizing the terms that are unique to only a few clusters, or terms whose cluster occurrence frequencies are highly skewed (that is, they occur with large occurrence frequencies in some clusters, and with much lower frequencies in many others). The second result is produced by deemphasizing terms that occur in many different clusters.

Two parameters may be introduced to be used in carrying out the required transformations [5]:

	NC(k)	the number of clusters in which term k occurs (a term occurs in a cluster if it is assigned to at least one document in that cluster);
and	CF(k,j)	the cluster frequency of term k in cluster j that is, the number of documents in cluster j in which term k occurs.

For a collection arranged into p clusters, the average cluster frequency $\overline{CF}(k)$ may then be defined from CF(k,j) as

$$CF(k) = \frac{1}{p} \sum_{j=1}^{p} CF(k,j).$$

Given the above parameters, the skewness of the occurrence frequencies of the terms may now be measured by a factor such as

$$F_1 = |\overline{CF}(k) - CF(k,j)|.$$

On the other hand, a factor F_2 inverse to $NC(k)$ (for example, $1/NC(k)$) can be used to reflect the rarity with which term k is assigned to the various clusters. By multiplying the weight of each term k in each cluster j by a factor proportional to $F_1 \cdot F_2$ a suitable spreading out should be obtained in the document space. Contrariwise, the space will be compressed when a multiplicative factor proportional to $1/F_1 \cdot F_2$ is used.

The output of Table 3 shows that a modification of term weights by the $F_1 \cdot F_2$ factor produces precisely the anticipated effect: the similarity between documents included in the same cluster (factor x) is now greater, whereas the similarity between different cluster centroids (factor y) has decreased. Overall, the space density measure (y/x) decreases by 18 and 11 percent respectively for the two cluster organizations. The average retrieval performance for the spread-out space shown at the bottom of Table 3 is improved by a few percentage points.

The corresponding results for the compression of the space using a transformation factor of $1/F_1 \cdot F_2$ are shown in Table 4. Here the similarity between documents inside a cluster decreases, whereas the similarity between cluster centroids increases. The overall space density measure (y/x) increases by 11 and 16 percent for the two cluster organizations compared with the space representing the standard term frequency weighting. This dense document space produces losses in recall and precision performance of 12 to 13 percent.

Taken together, the results of Tables 1 to 4 indicate that retrieval performance and document space density appear inversely related, in the sense that effective (questionable) indexing methods in terms of recall and precision are associated with separated (compressed) document spaces; on the other hand, artificially generated alterations in the space densities appear to produce the anticipated changes in performance.

The foregoing evidence thus confirms the usefulness of the "term discrimination" model and of the automatic indexing theory based on it. These questions are examined briefly in the remainder of this study.

4. The Discrimination Value Model

For some years, a document indexing model known as the term discrimination model has been used experimentally. [2,6] This model bases the value of an index term on its "discrimination value" DV, that is, on an index which measures the extent to which a given term is able to increase the differences among document vectors when assigned as an index term to a given collection of documents. A "good" index term — one with a high discrimination value — decreases the similarity between documents when assigned to the collection, as shown in the example of Fig. 5. The reverse obtains for the "bad" index term with a low discrimination value.

To measure the discrimination value of a term, it is sufficient to take the difference in the space densities before and after assignment of the particular term. Specifically, let the density of the complete space be measured by a function Q such as that of equation (3); that is, by the sum of the similarities between all documents and the space centroid. The contribution of a given term k to the space density may be ascertained by computing the function

$$DV_k = Q_k - Q, \tag{4}$$

where Q_k is the compactness of the document space with term k deleted from all document vectors. If term k is a good discriminator, valuable for content identification, then $Q_k > Q$, that is, the document space after removal of term k will be more compact (because upon assignment of that term to the documents of a collection the documents will resemble each other less and the space spreads out). Thus for good discriminators $Q_k - Q > 0$; the reverse obtains for poor discriminators for which $Q_k - Q < 0$.

Because of the manner in which the discimination values are defined, it is clear that the good discriminators must be those with uneven occurrence frequency distributions which cause the space to spread out when assigned by decreasing the similarity between the individual documents. The reverse is true for the bad discriminators. A typical list including the ten best terms and the ten worst terms in discrimination value order (in order by the $Q_k - Q$ value) is shown in Table 5 for a collection of 425 articles in world affairs from Time magazine. A total of 7569 terms are used for this collection, exclusive of the common English function words that have been deleted.

In order to translate the discrimination value model into a possible theory of indexing, it is necessary to examine the properties of good and bad discriminators in greater detail. Fig. 6 is a graph of the terms assigned to a sample collection of 450 documents in medicine, presented in order by their document frequencies. For each class of terms — those of document frequency 1, document frequency 2, etc. ... — the average rank of the corresponding terms is given in discrimination value order (rank 1 is assigned to the best discriminator and rank 4726 to the worst term for the 4726 terms of the medical collection).

Fig. 6 shows that terms of low document frequency — those that occur in only one, or two, or three documents — have rather poor average discrimination ranks. The several thousand terms of document frequency 1 have an average rank exceeding 3000 out of 4726 in discrimination value order. The terms with very high document frequency — at least one term in the medical collection occurs in as many as 138 documents out of 450 — are even worse discriminators; the terms with document frequency greater than 25 have average discrimination values in excess of 4000 in the medical collection. The best discriminators are those whose document frequency is neither too low nor too high.

The situation relating document frequency to term discrimination value is summarized in Fig. 7. The 4 percent of the terms with the highest document frequency, representing about 50 percent of the total term assignments to the documents of a collection, are the worst discriminators. The 70 percent of the terms with the lowest document frequency are generally poor discriminators. The best discriminators are the 25 percent whose document frequency lies approximately between n/100 and n/10 for n documents.

If the model of Fig. 7 is a correct representation of the situation relating to term importance, the following indexing strategy results [6,7]:

a) Terms with medium document frequency should be used for content identification directly, without further transformation.

b) Terms with very high document freauency should be moved to the left on the document frequency spectrum by transforming them into entities of lower frequency; the best way of doing this is by taking high-frequency terms and using them as components of indexing phrases — a phrase such as "programming language" will necessarily exhibit lower document frequency than either "program", or "language" alone.

c) Terms with very low document frequency should be moved to the right on the document frequency spectrum by being transformed into entities of higher frequency; one way of doing this is by collecting several low frequency terms that appear semantically similar and including them in a common term (thesaurus) class. Each thesaurus class necessarily exhibits a higher document frequency than any of the component members that it replaces.

The indexing theory which consists in using certain elements extracted from document texts directly as index terms, combined with phrases made up of high frequency components and thesaurus classes defined from low frequency elements has been tested using document collections in aerodynamics (CRAN), medicine (MED), and world affairs (TIME). [2,6,7] A typical recall-precision plot showing the effect of the right-to-left phrase transformation is shown in Fig. 8 for the Medlars collection of 450 medical documents. When recall is plotted against precision, the curve closest to the upper right-hand corner of the graph (where both recall and precision are close to 1) reflects the best performance. It may be seen from Fig. 8 that the replacement of the high-frequency nondiscriminators by lower frequency phrases improves the retrieval performance by an average of 39 percent (the precision values at the ten fixed recall points are greater by an average of 39 percent).

The performance of the right-to-left (phrase) transformation and left-to-right (thesaurus) transformation is summarized in Table 6 for the three previously mentioned test collections. The precision values obtainable are near 90 percent for low recall, between 40 and 70 percent for medium recall, and between 15 and 45 percent at the low recall end of the performance spectrum. The overall improvement obtainable by phrase and thesaurus class assignments over the standard term frequency process using only the unmodified, single terms ranges from 17 percent for the world affairs collection to 50 percent for the medical collection.

A conclusive proof relating the space density analysis and the resulting document frequency indexing model to optimality in the retrieval performance cannot be furnished. However, the model appears to perform well for collections in several different subject areas, and the performance results produced by applying the theory have not in the authors' experience been surpassed by any other manual or automatic indexing and analysis procedures tried in earlier experiments. The model may then lead to the best performance obtainable with ordinary document collections operating in actual user environments.

References

[1] G. Salton, Automatic Information Organization and Retrieval, McGraw Hill Book Co., New York, 1968, chapter 4.

[2] 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.

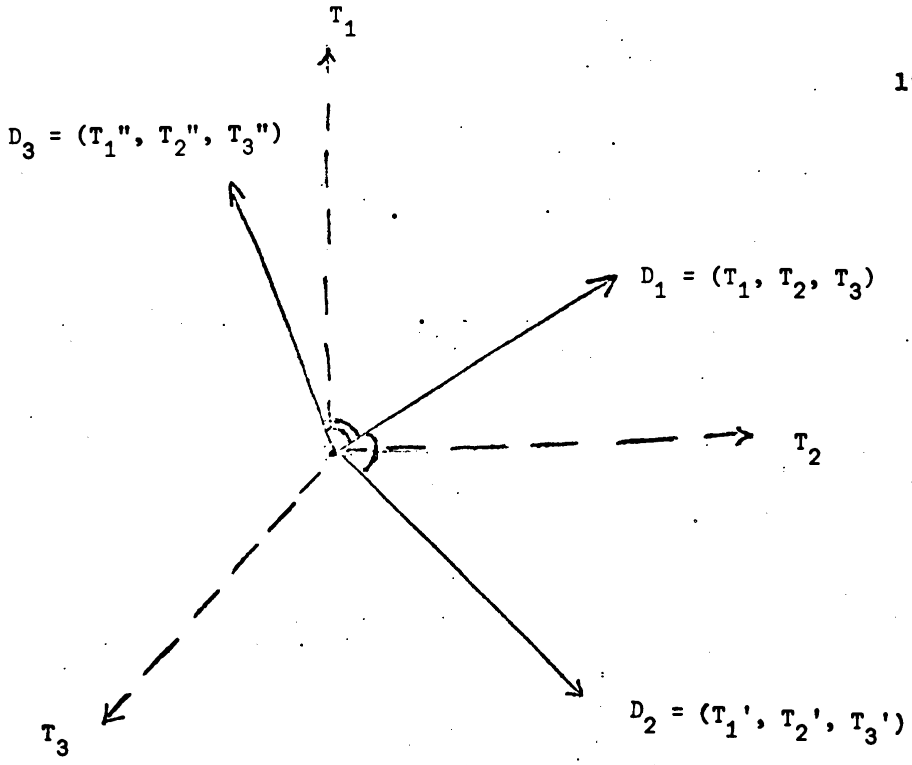
[3] K. Sparck Jones, A Statistical Interpretation of Term Specificity and its Application to Retrieval, Journal of Documentation, Vol. 28, No. 1, March 1972, p. 11-20.

[4] R.E. Williamson, Real-time Document Retrieval, Cornell University Ph.D. Thesis, Department of Computer Science, June 1974.

[5] A. Wong, An Investigation of the Effects of Different Indexing Methods on the Document Space Configuration, Scientific Report No. ISR-22, Department of Computer Science, Cornell University, to appear.

[6] G. Salton, A Theory of Indexing, Technical Report No. TR 74-203, Department of Computer Science, Cornell University, March 1974.

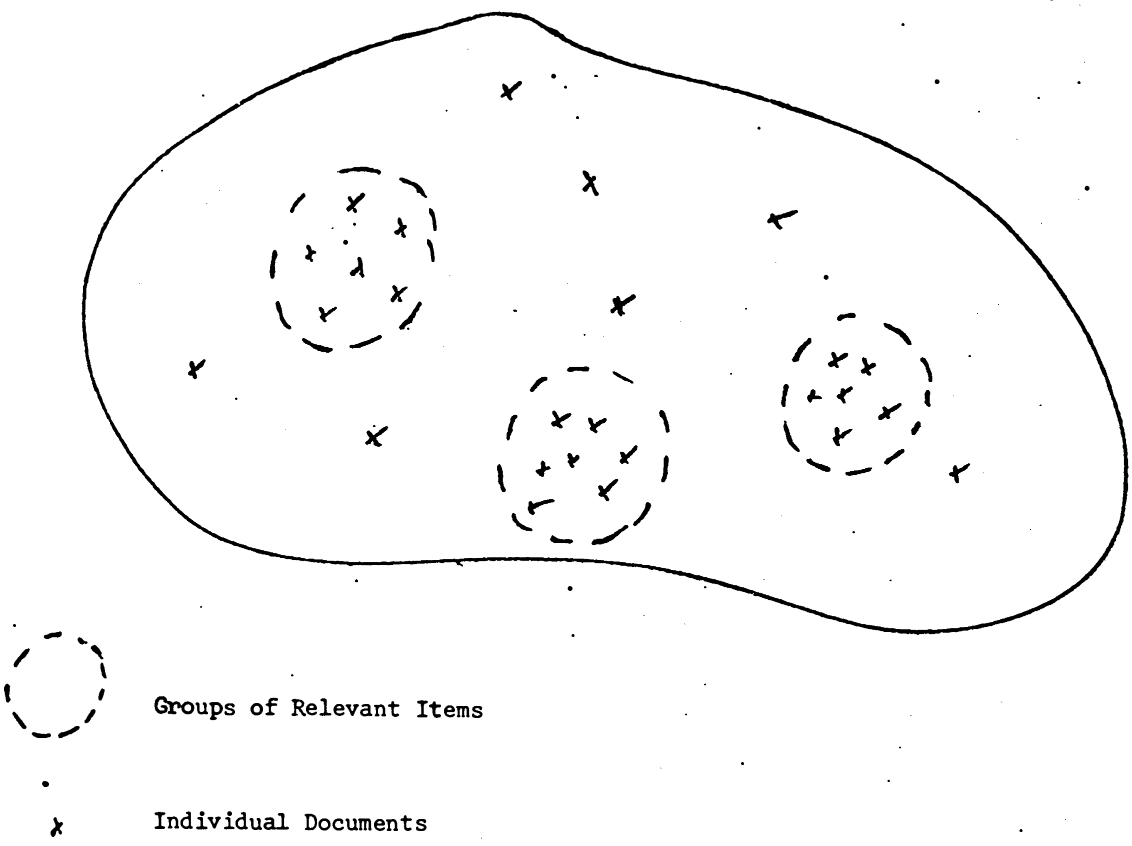
[7] G. Salton, C.S. Yang, and C.T. Yu, Contribution to the Theory of Indexing, Proc. IFIP Congress-74, Stockholm, August 1974.



Vector Representation of

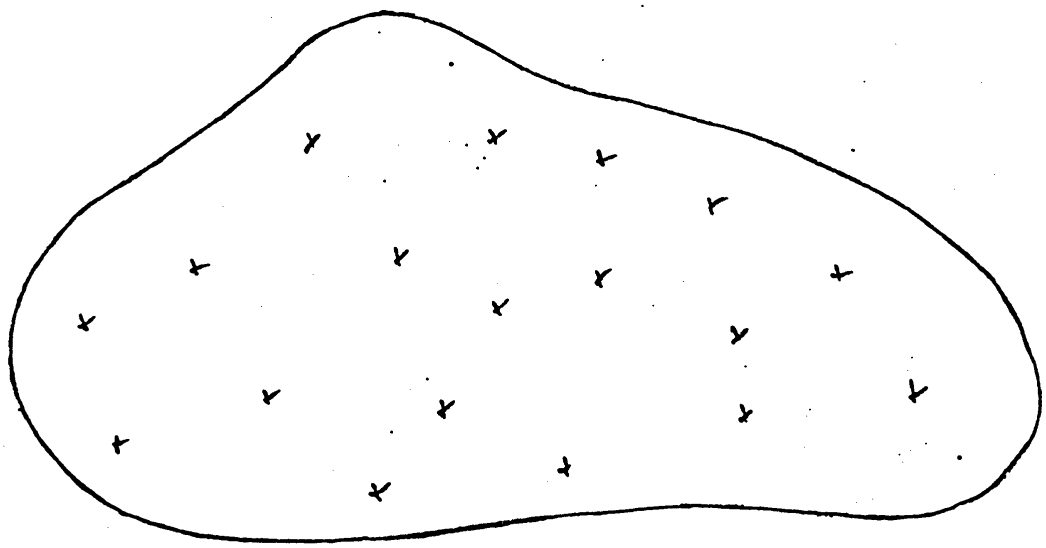
Document Space

Fig. 1



Ideal Document Space

Fig. 2

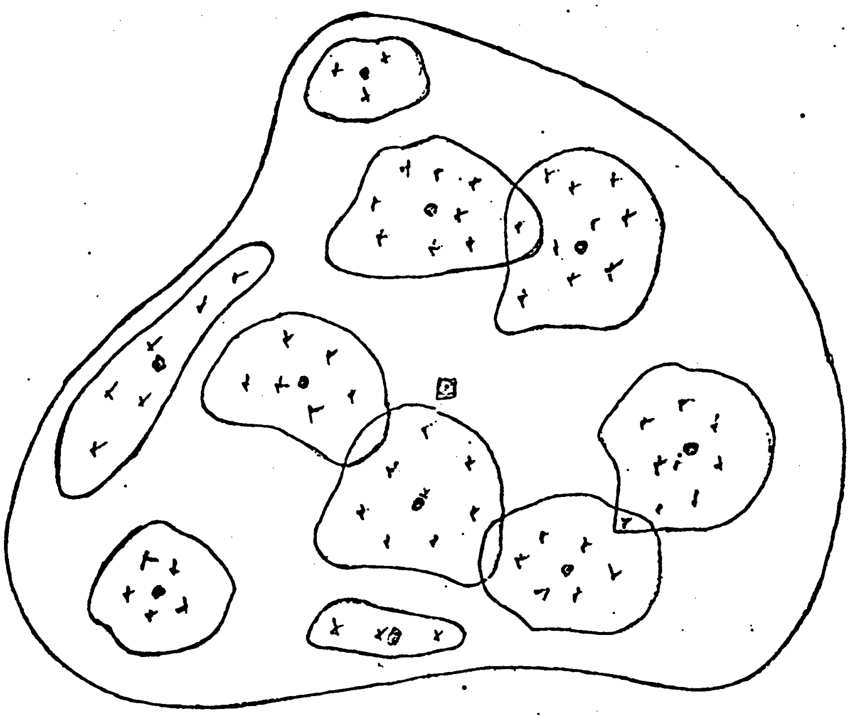


X Individual Document

Space with Maximum Separation

Between Document Pairs

Fig. 3

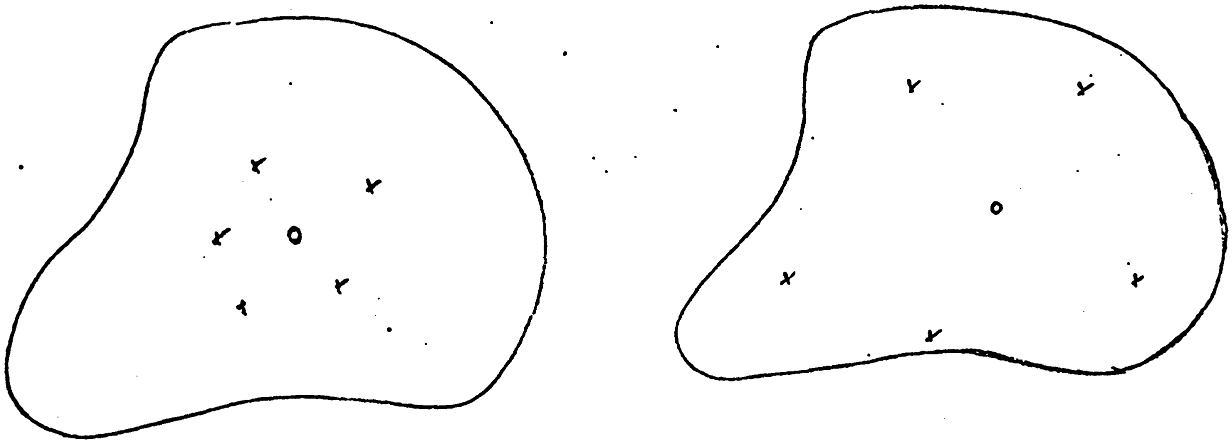


Cluster Centroid

Main Centroid

Clustered Document Space

Fig. 4



Before Assignment of Term

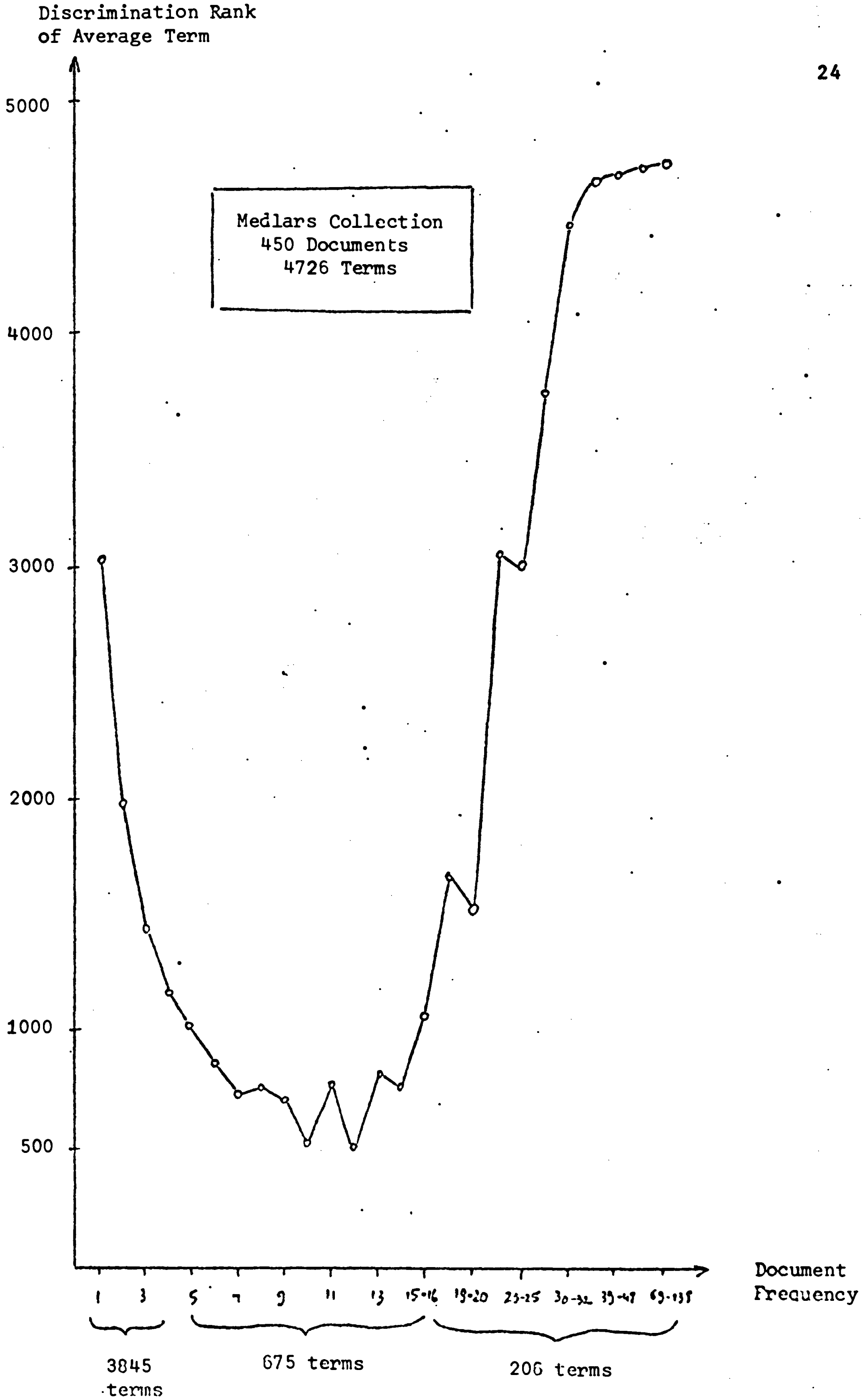
After Assignment of Term

X Document

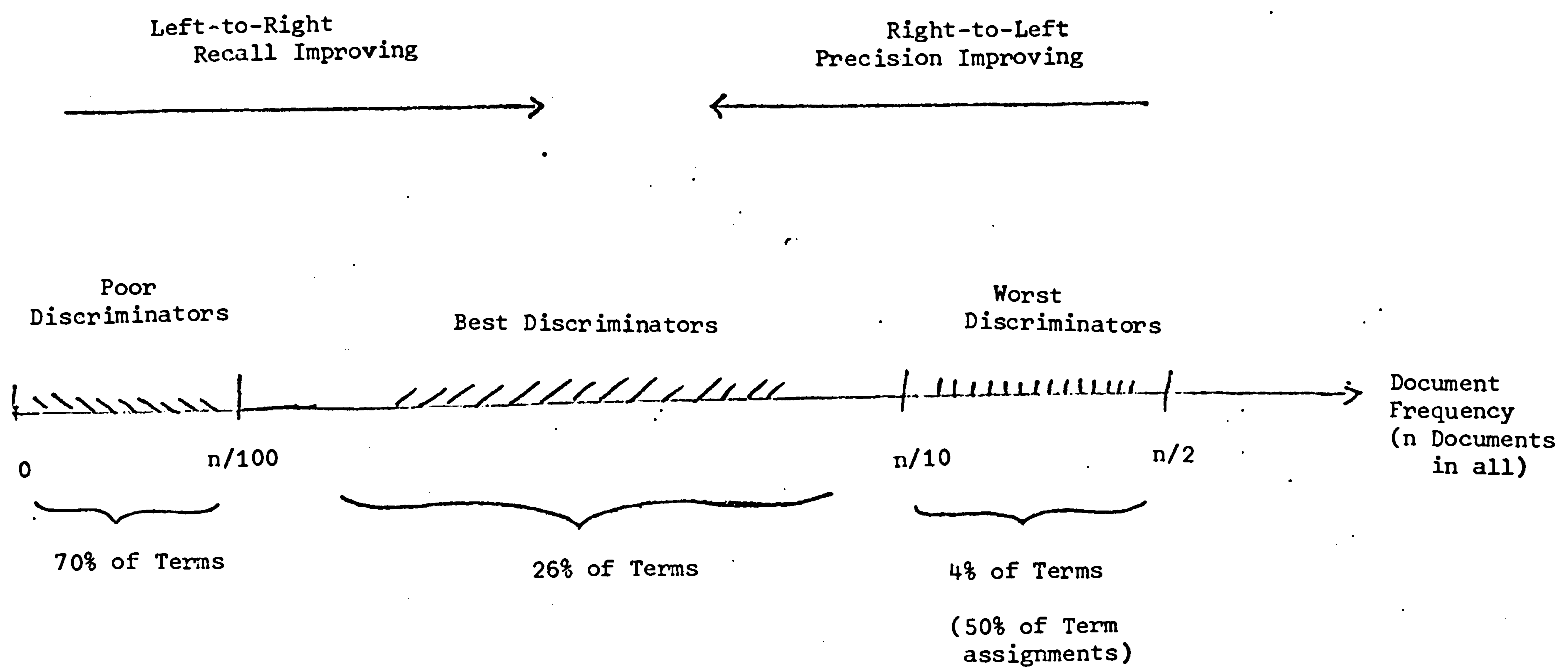
o Main Centroid

Operation of Good Discriminating Term

Fig. 5



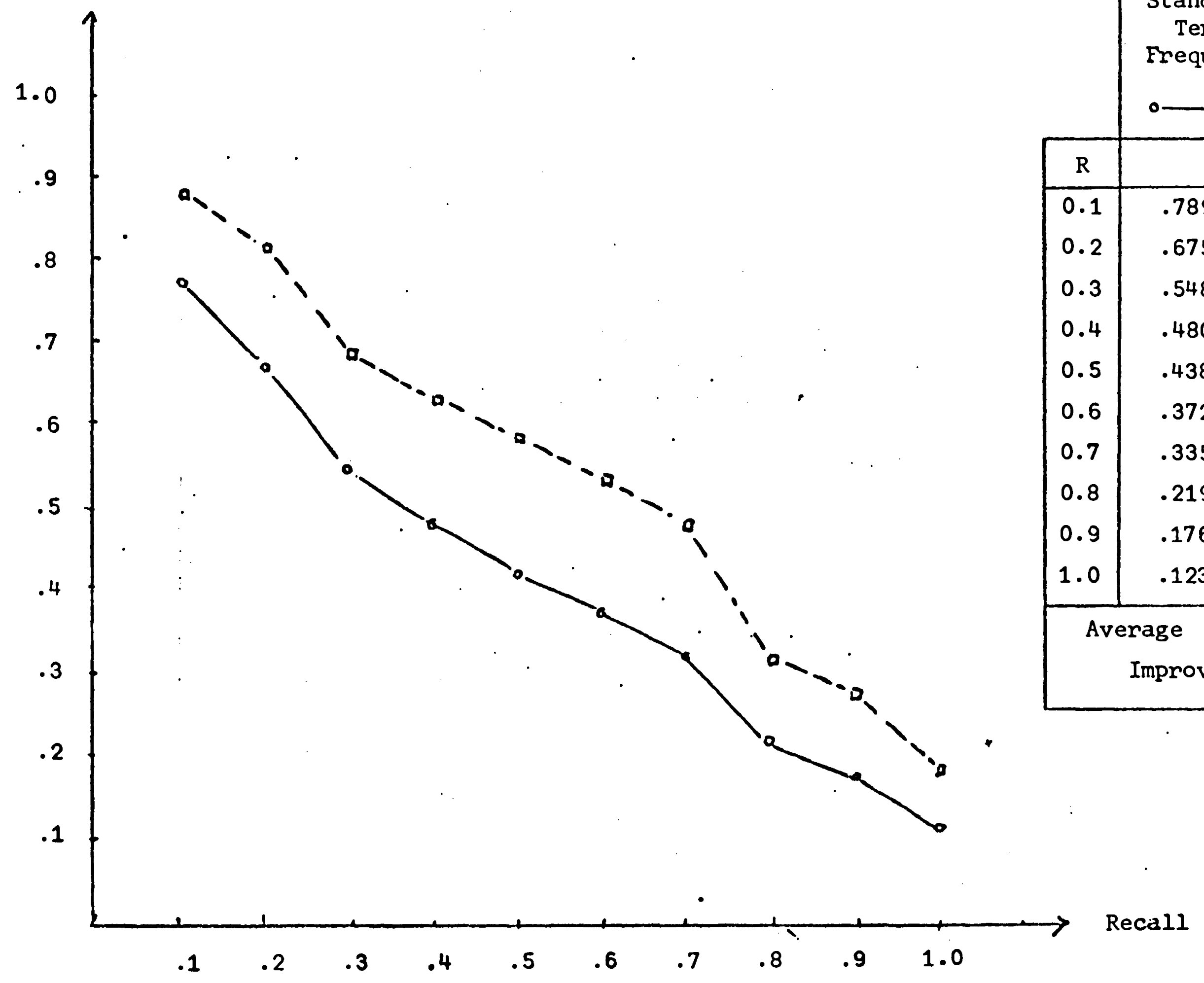
Average Discrimination Value Rank of Terms



Summarization of Discrimination Value of

Terms in Frequency Ranges

Fig. 7



R	Standard Term Frequency (o———o) Precision	Phrase Assignment (□———□) Precision
0.1	.7891	.8811
0.2	.6750	.8149
0.3	.5481	.6992
0.4	.4807	.6481
0.5	.4384	.5930
0.6	.3721	.5450
0.7	.3357	.4867
0.8	.2195	.3263
0.9	.1768	.2767
1.0	.1230	.1969
Average Improvement		+ 39%

Average Recall-Precision Comparison for Phrases

	Cluster Organization A (155 clusters; 2.1 overlap)		Cluster Organization B (83 clusters; 1.3 overlap)	
Type of Indexing	Standard Term Frequency Weights (f_i^k)	Term - Frequency with Inverse Doc. Freq. $(f_i^k \cdot IDF_k)$	Standard Term Frequency Weights (f_i^k)	Term - Frequency with Inverse Doc. Freq. $(f_i^k \cdot IDF_k)$
Recall-Precision Output (J. Doc. December 73)	—	+14%	—	+14%
Average Similarity Between Documents and Corresponding Cluster Centroids (x)	.712	.668 (-.044)	.650	.589 (-.061)
Average Similarity Between Cluster Centroids and Main Centroid	.500	.454 (-.046)	.537	.492 (-.045)
Average Similarity Between Pairs of Cluster Centroids (y)	.273	.209 (-.046)	.315	.252 (-.063)
Ratio y/x	$\frac{.273}{.712} = .383$	$\frac{.209}{.668} = .318$ (-19%)	$\frac{.315}{.650} = .485$	$\frac{.252}{.589} = .428$ (-12%)

Effect of Performance Improvement on Space Density

Table 1

Type of Indexing	Cluster Organization A (155 clusters; 2.1 overlap): Standard Term Frequency Weights (f_i^k)	Cluster Organization A (155 clusters; 2.1 overlap): Term Frequency with Document Frequency $(f_i^k \cdot DF_k)$	Cluster Organization B (83 clusters; 1.3 overlap): Standard Term Frequency Weights (f_i^k)	Cluster Organization B (83 clusters; 1.3 overlap): Term Frequency with Document Frequency $(f_i^k \cdot DF_k)$
Recall- Precision Output	—	-10.1%	—	-10.1%
Average Similarity Between Documents and Corresponding Cluster Centroids (x)	.712	.741 (+.029)	.650	.696 (+046)
Average Similarity Between Cluster Centroids and Main Centroid	.500	.555 (+.055)	.537	.574 (+.037)
Average Similarity Between Pairs of Cluster Centroids (y)	.273	.329 (+.056)	.315	.362 (+.047)
Ratio y/x	$\frac{.273}{.712} = .383$	$\frac{.329}{.741} = .444$ (+16%)	$\frac{.315}{.650} = .485$	$\frac{.362}{.696} = .520$ (+7%)

Effect of Performance Deterioration on Space Density

Table 2

	Cluster Organization A (155 clusters; 2.1 overlap)		Cluster Organization B (83 clusters; 1.3 overlap)	
	Standard Cluster Density (term frequency weights)	High Cluster Density (emphasis of low frequency and skewed terms)	Standard Cluster Density (term frequency weights)	High Cluster Density (emphasis of low frequency and skewed terms)
verage Similarity etween Documents nd their Centroids (x)	.712	.730 (+.018)	.650	.653 (+.003)
verage Similarity etween Cluster entroids and ain Centroid	.500	.477 (-.023)	.537	.528 (-.009)
verage Similarity etween Pairs f Centroids (y)	.273	.229 (-.044)	.315	.281 (-.034)
atio y/x	$\frac{.273}{.712}$ = .383	$\frac{.229}{.730}$ = .314 (-18%)	$\frac{.315}{.650}$ = .485	$\frac{.281}{.653}$ = .430 (-11%)
ecall-Precision Comparison	—	+2.6%	—	+2.3%

Effect of Low Cluster Density on Performance

Table 3

	Cluster Organization A (155 clusters; 2.1 overlap)		Cluster Organization B (83 clusters; 1.3 overlap)	
	Standard Cluster Density (term frequency weights)	High Cluster Density (emphasis on high frequency and even terms)	Standard Cluster Density (term frequency weights)	High Cluster Densi (emphasis on high frequen and even ter
Average Similarity between Documents and their Centroids (x)	.712	.681 (-.031)	.650	.645 (-.005)
Average Similarity between Cluster Centroids and Main Centroid	.500	.523 (+.023)	.537	.531 (-.006)
Average Similarity between Pairs of Centroids (y)	.273	.290 (+.017)	.315	.364 (+.049)
Ratio y/x	$\frac{.273}{.712} = .383$	$\frac{.290}{.681} = .426$ (+11%)	$\frac{.315}{.650} = .485$	$\frac{.364}{.645} = .561$ (+16
Recall-Precision Comparison	—	-12.4%	—	-13.3%

Effect of High Cluster Density on Performance

Table 4

Good Terms		Poor Terms	
1.	Buddhist	7560.	Work
2.	Diem	7561.	Lead
3.	Lao	7562.	Red
4.	Arab	7563.	Minister
5.	Viet	7564.	Nation
6.	Kurd	7565.	Party
7.	Wilson	7566.	Commune
8.	Baath	7567.	U.S.
9.	Park	7568.	Govern
10.	Nenni	7569.	New

Terms in Discrimination Value Order

(1963 Time Magazine)

Table 5

CRAN 424	MED 450	TIME 425
Automatic Phrases vs. Standard Term Frequency +32%	Automatic Phrases vs. Standard Term Frequency +39%	Automatic Phrases vs. Standard Term Frequency
Automatic Phrases Plus Thesaurus vs. Standard Run +33%	Automatic Phrases Plus Thesaurus vs. Standard Run +50%	Automatic Phrases Plus Thesaurus vs. Standard Run
Best Precision Low Recall 0.89 Medium Recall 0.43 High Recall 0.13	Best Precision Low Recall 0.88 Medium Recall 0.61 High Recall 0.23	Best Precision Low Recall 0.85 Medium Recall 0.70 High Recall 0.45

Summary of Recall-Precision Evaluation (Three Collections)

Table 6