A THEORY OF TERM IMPORTANCE IN AUTOMATIC TEXT ANALYSIS

G. Salton*, C.S. Yang*, and C.T. Yu+

TR 74-208

July 1974

Department of Computer Science
Cornell University
Ithaca, New York 14853

* Department of Computer Science, Cornell University, Ithaca, N.Y. 1485(

* Department of Computer Science, University of Alberta, Edmonton, Albe

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.

The theory is validated by citing experimental results.

1. Document Space Configuration

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

$$D_i = (w_{i1}, w_{i2}, \ldots, w_{it}) \quad (1)$$

where w_{ij} represents the weight of term j in the vector corresponding to the ith document. Given two documents D_i and D_j, it is possible to define a measure of relatedness $s(D_i, D_j)$ 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.

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 ith vector element c_i of the centroid can simply by defined as the average of the ith 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. [1,2] 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

$$Q = \sum_{i=1}^{n} s(C, D_i) \qquad (2)$$

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 kth term removed from all document (and from the centroid) vectors. The discriminatio

* 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

$$DV_k = Q_k - Q. \tag{3}$$

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 $Q_k > Q$ 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 aerodynamics (Cranfield collection, 424 documents comprising 2651 distinct terms), medicine (Medlars 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,

$$w_{kj} = f_{kj} \cdot DV_k, \tag{4}$$

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 d_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 $d_k = I$. 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, which 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 ranks 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 3200 terms out of 4700 with document frequencies of 1 or 2, and in the Time collection to 9900 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 3800 Medlars terms, or 11300 Time terms, ranging in document frequency from 1 to 6, 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 high 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 46 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 1, out of the top 50, or 11 of the top 100 are included among the first 8916 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 n/100 and n/10 — then the following term transformations should be undertaken:

a) Terms whose document frequency lies between n/100 and 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 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 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 <u>indexing phrases</u>. 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_{ip} and w_{iq}, 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_{ipq} defined as

$$w_{ipq} = \frac{w_{ip} + w_{iq}}{2} . \qquad (5)$$

A somewhat more refined weighting method uses w_{ipq} 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 [4]:

$$IDF_k = \lceil \log_2 n \rceil - \lceil \log_2 d_k \rceil + 1,$$

where 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 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:

$$IDF_{pq} = (\log n - \frac{\log d_p + \log d_q}{2}), \qquad (6)$$

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 $\underline{W}_{ipq}$ 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)):

$$\underline{W}_{ipq} = \left[\log n - \frac{\log d_p + \log d_q}{2}\right] \cdot \left[\frac{w_{ip} + w_{iq}}{2}\right] .* \qquad (7)$$

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 the 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

$$D_{pq} = d_p + d_q - d_{pq}$$

term q, and both p and q, respectively. In general D_{pq} may be expected to be larger than either d_p or d_q individually. When m 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 assigned 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 relatively low overall document frequency in the collection.

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. [5,6,7,8] 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_{ki}) 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 24 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, 0.40, and 0.20 at high, medium, and low precision are transformed into average figures of 0.90, 0.60 and 0.30 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

[1] 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.

[2] G. Salton, A. Wong, and C.S. Yang, A Vector Space Model for Automatic Indexing, Technical Report No. 74-208, Department of Computer Science, Cornell University, Ithaca, N.Y., July 1974.

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

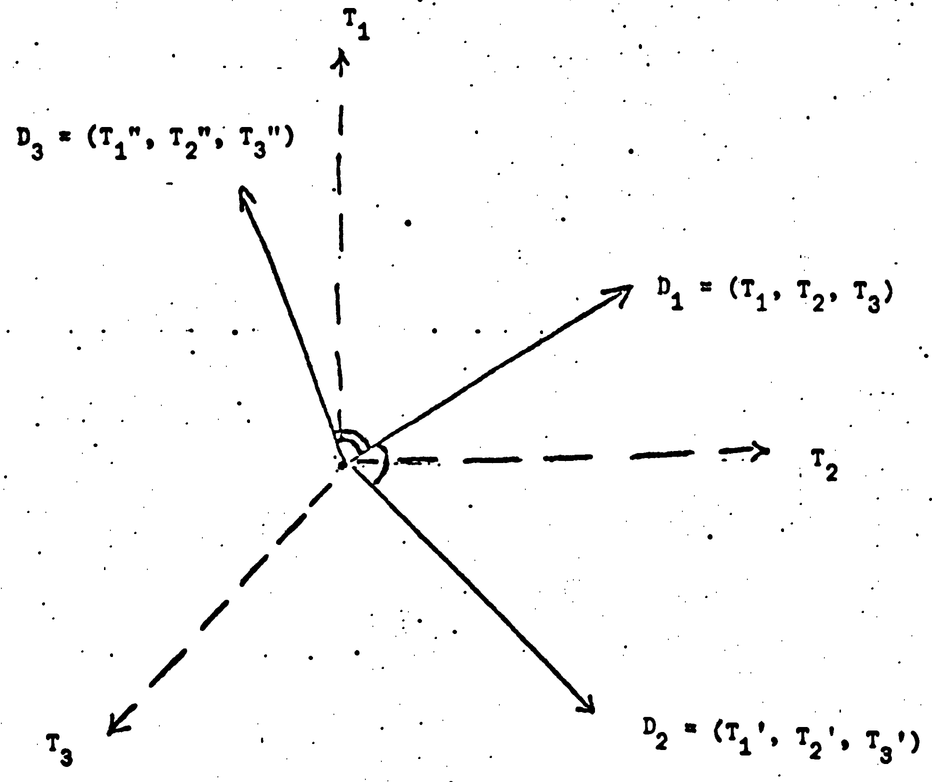
[4] K. Sparck Jones, A Statistical Interpretation of Term Specificity and its Application to Retrieval, Journal of Documentation, Vol. 28, No. 1, March 1972, p. 11-20.

[5] K. Sparck Jones, Automatic Keyword Classifications, Butterworths, London, 1971.

[6] C.C. Gotlieb and S. Kumar, Semantic Clustering of Index Terms, ACM Journal, Vol. 15, No. 4, October 1968, p. 493-513.

[7] G. Salton, Experiments in Automatic Thesaurus Construction for Information Retrieval, Information Processing 71, North Holland Publishing Co., Amsterdam, 1972, p. 115-123.

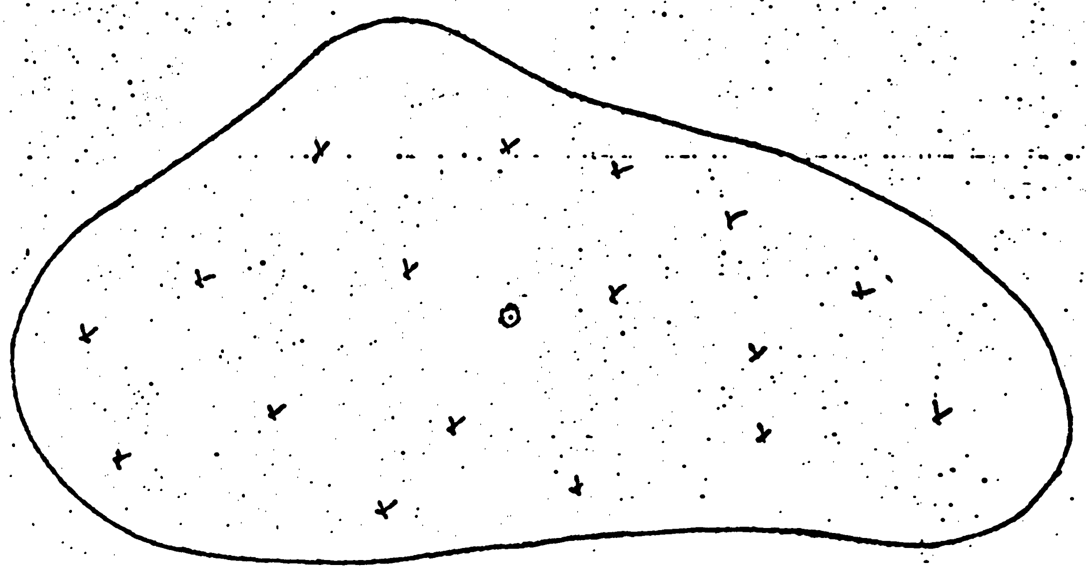
[8] G. Salton, C.S. Yang, and C.T. Yu, Contributions to the Theory of Indexing, Proc. IFIP Congress-74, Stockholm, August 1974.



Vector Representation of

Document Space

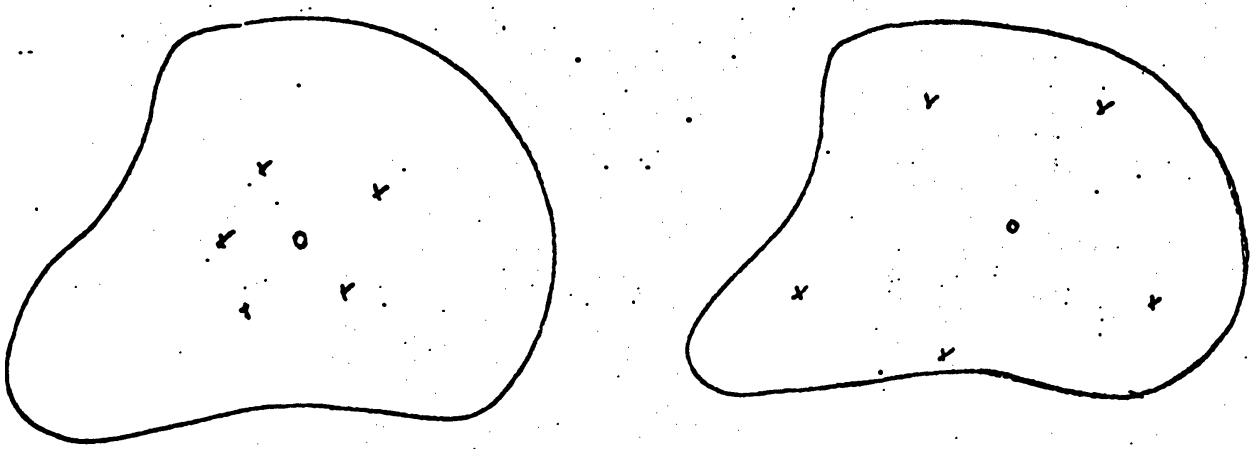
Fig. 1



O Centroid of Space
X Individual Document

Multidimensional Document Space

Fig. 2



Before Assignment of Term

After Assignment of Term

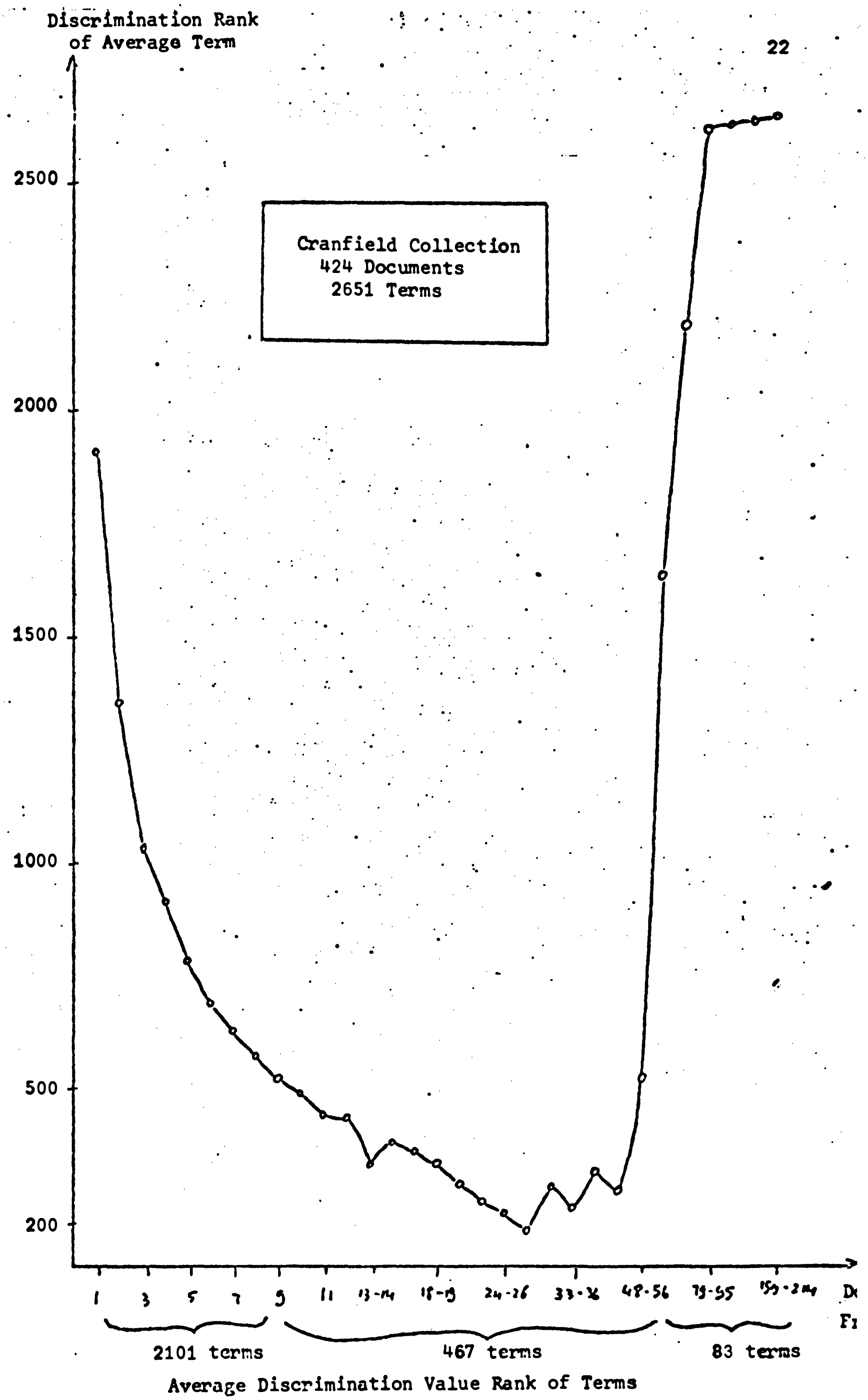
X Document

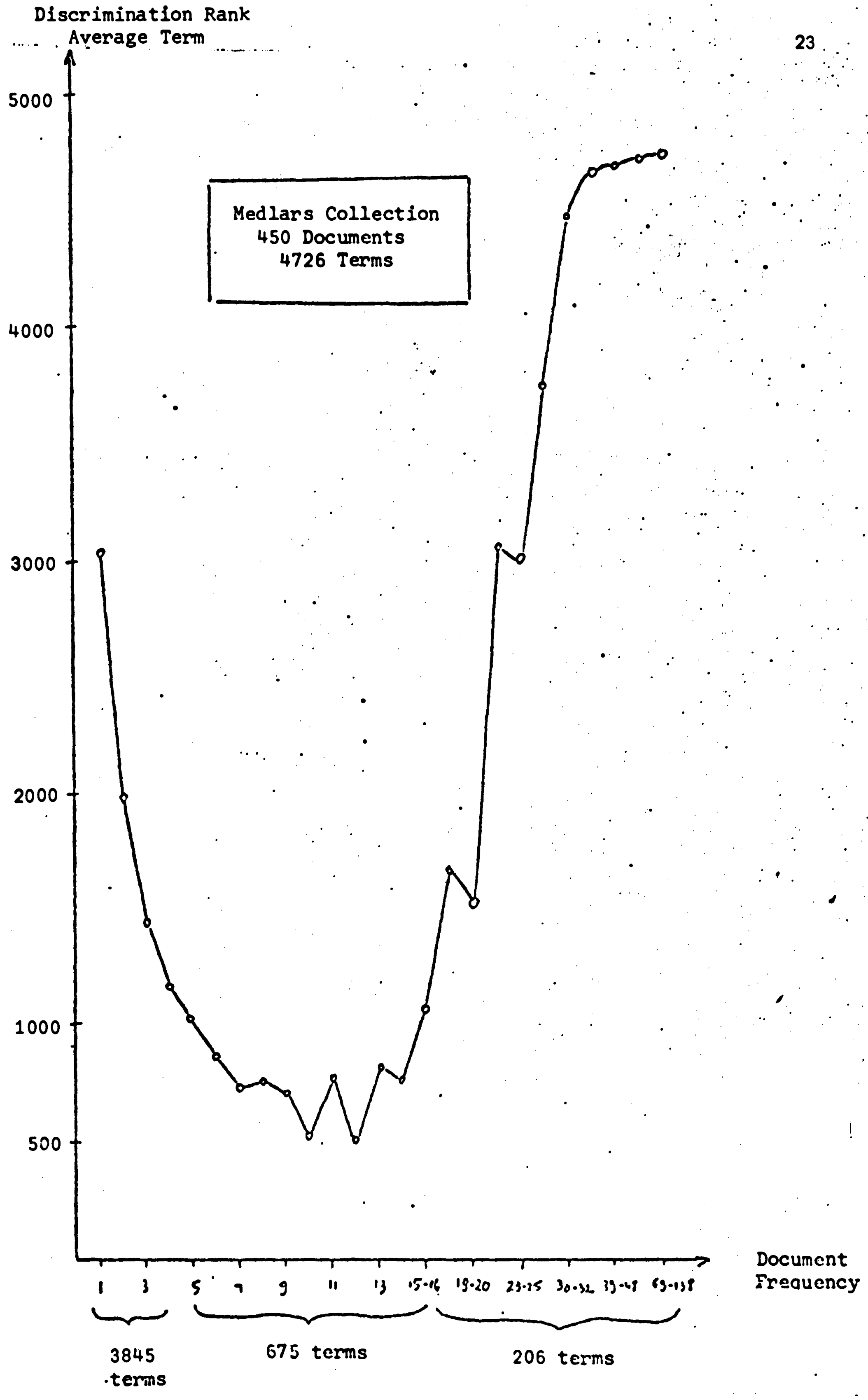
o Main Centroid

Operation of

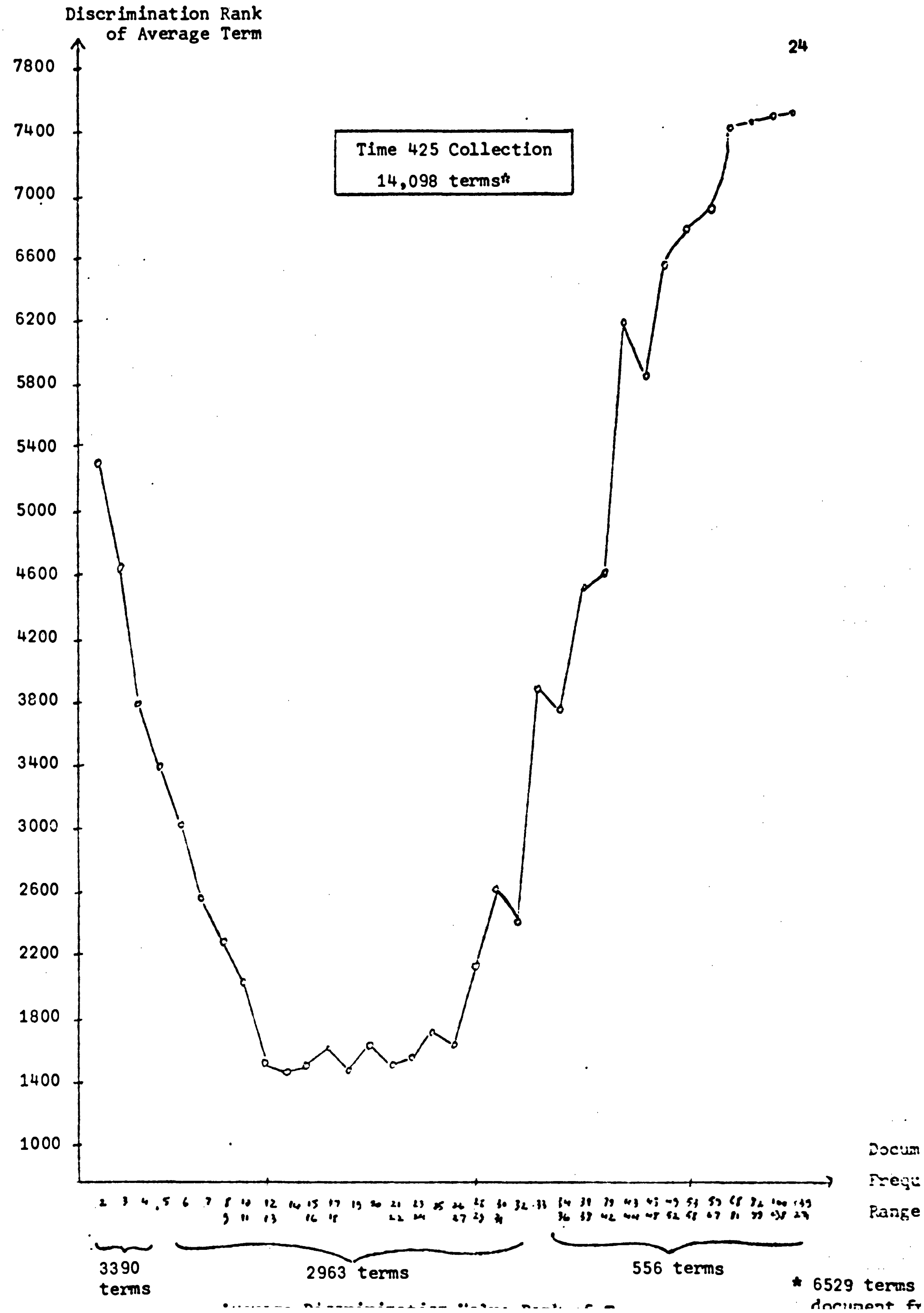
Good Discriminating Term

Fig. 3

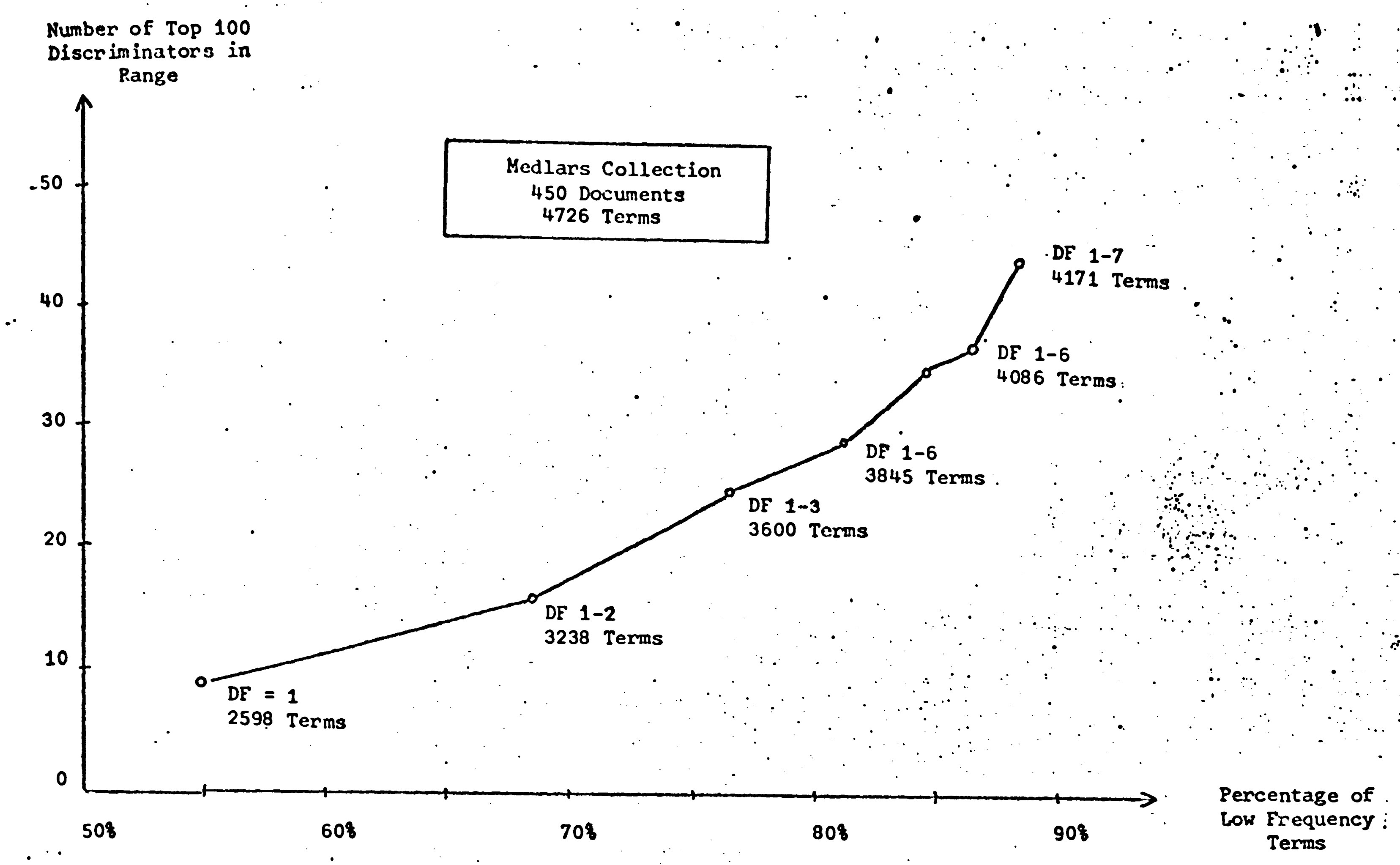

Discrimination Rank
of Average Term
2500
2000
1500
1000
500
200
Cranfield Collection
424 Documents
2651 Terms
1
3
5
7
9
11
13-14
18-19
24-26
33-36
48-56
2101 terms
467 terms
83 terms
Average Discrimination Value Rank of Terms



Average Discrimination Value Rank of Terms



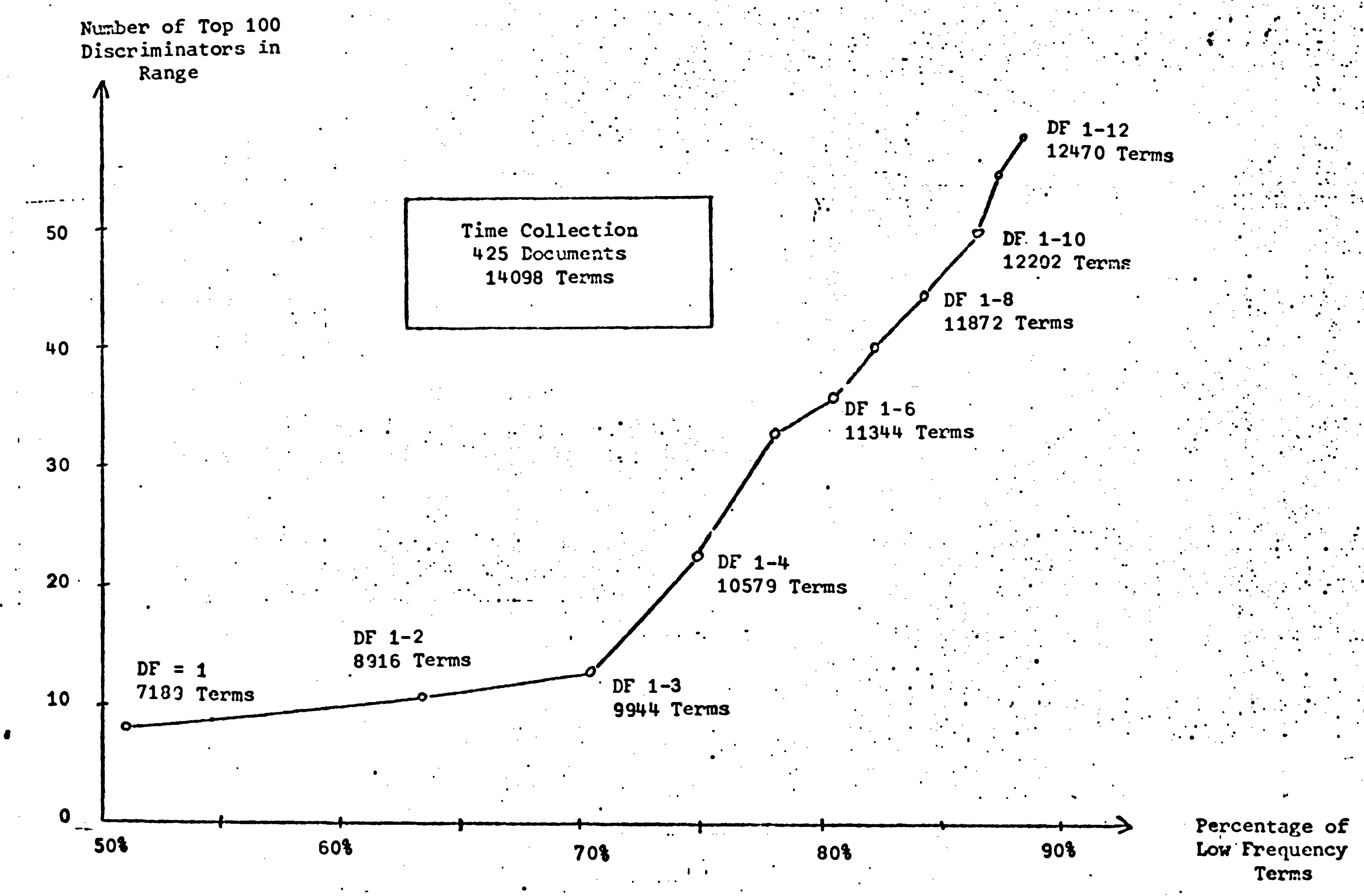
Discrimination Rank
of Average Term
7800
7400
7000
6600
6200
5800
5400
5000
4600
4200
3800
3400
3000
2600
2200
1800
1400
1000
Time 425 Collection
14,098 terms*
3390
terms
2963 terms
556 terms
* 6529 terms



Number of Good Discriminators Among

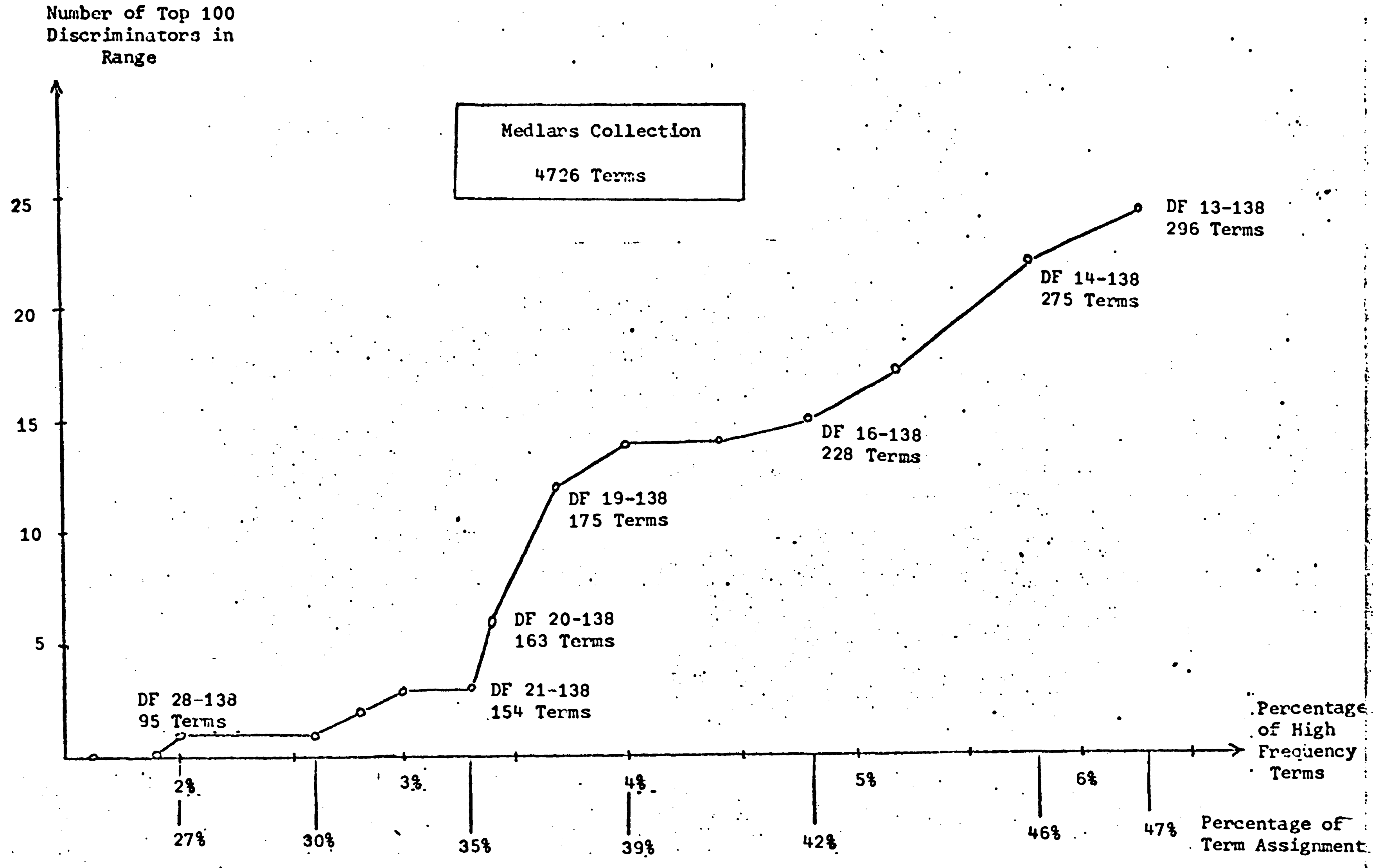
Low Frequency Terms

Fig. 5(a)



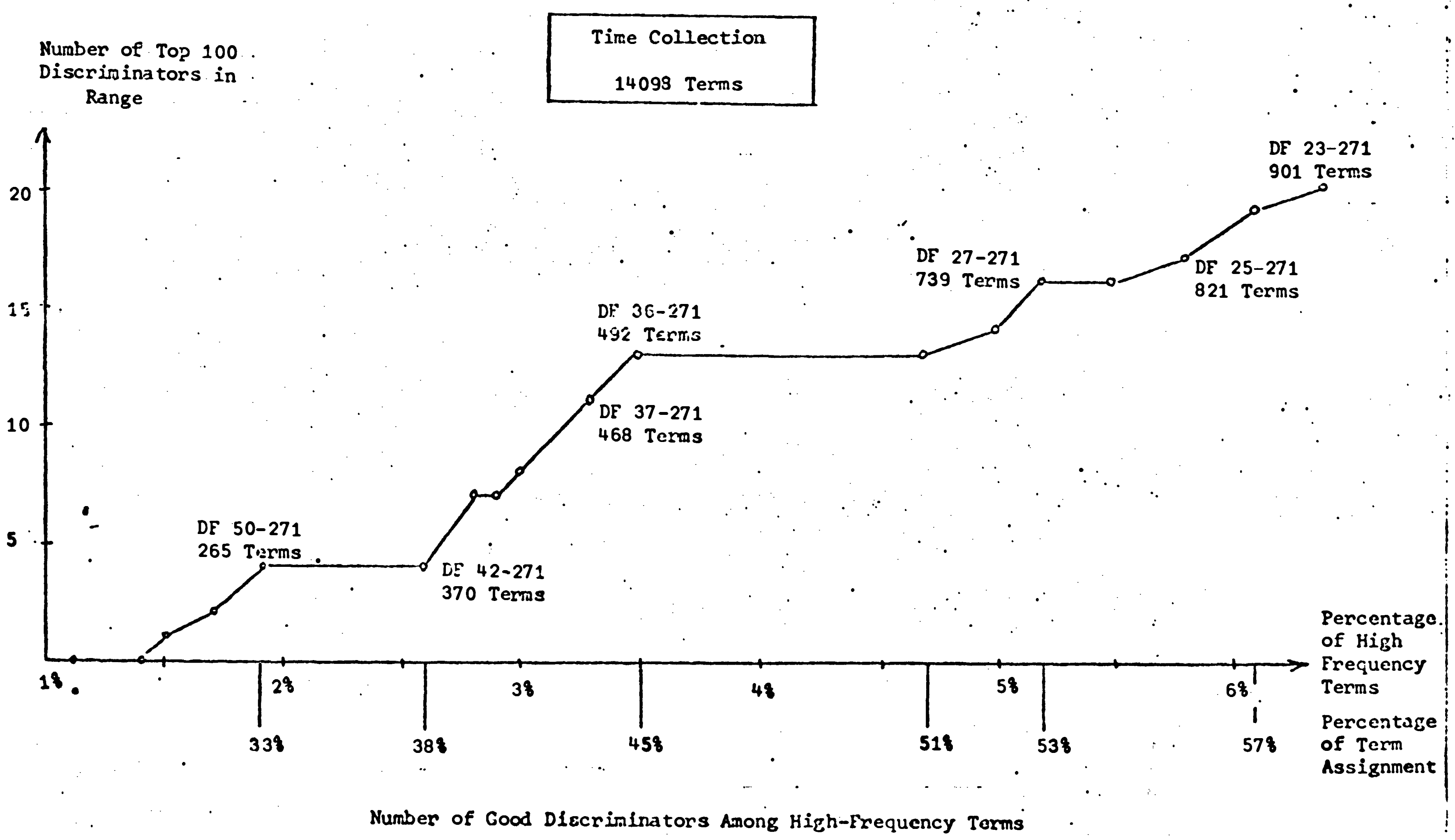
Number of Good Discriminators Among Low Frequency Terms

Fig. 5(b)

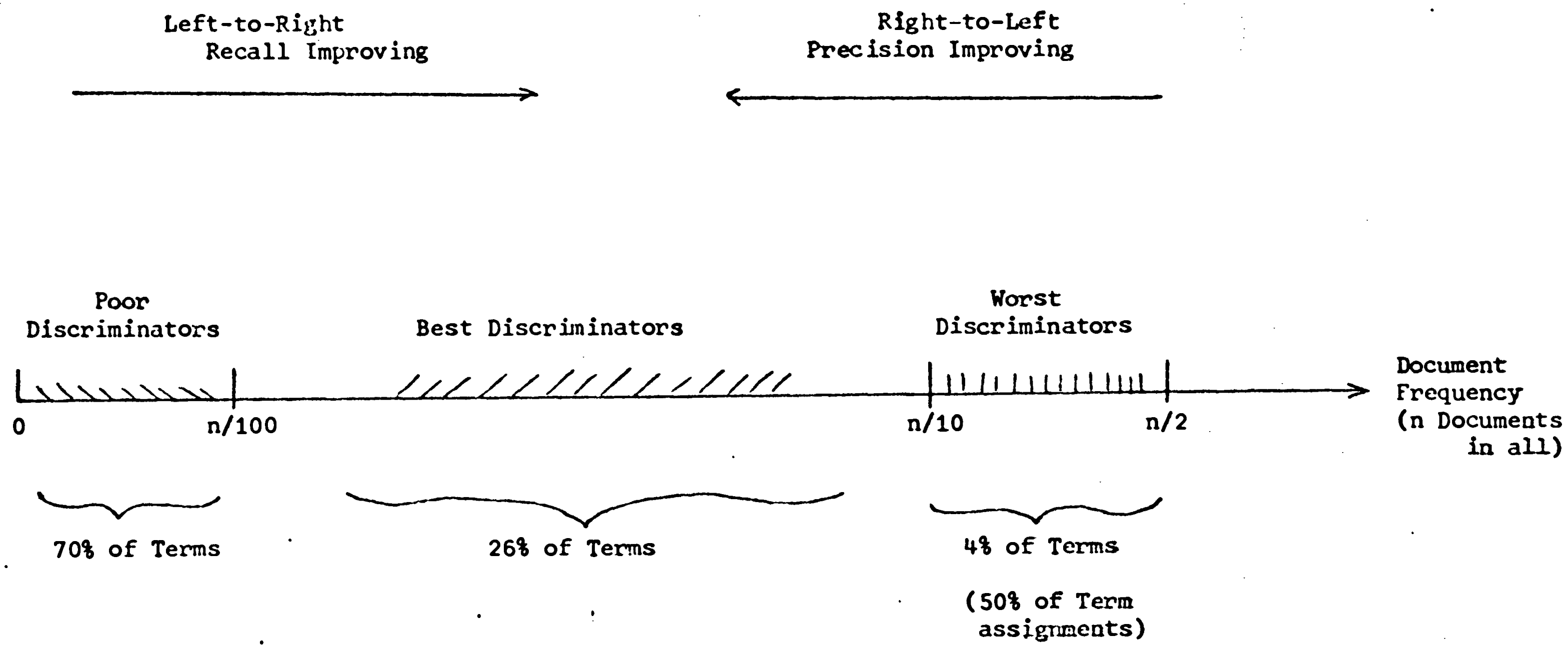


Number of Good Discriminators Among High Frequency Terms

Fig. 6(a)



Time Collection
14098 Terms
Number of Top 100 Discriminators in Range
20
15
10
5
DF 50-271
265 Terms
DF 42-271
370 Terms
DF 36-271
492 Terms
DF 37-271
468 Terms
DF 27-271
739 Terms
DF 25-271
821 Terms
DF 23-271
901 Terms
1%
2%
3%
4%
5%
6%
Percentage of High Frequency Terms
33%
38%
45%
51%
53%
57%
Percentage of Term Assignment
Number of Good Discriminators Among High-Frequency Terms



Summarization of Discrimination Value of

Terms in Frequency Ranges

Fig. 7

QUERY: COALITION GOVERNMENT TO BE FORMED IN ITALY BY THE LEFT-WING SOCIALISTS, THE REPUBLICANS, SOCIAL DEMOCRATS, AND CHRISTIAN DEMOCRATS.

DELETE COMMON WORDS AND ELIMINATE SUFFIXES:

COALI GOVERN FORM ITALY LEFT-W SOCIAL REPUBLIC
SOCIAL DEMOCRAT CHRIST DEMOCRAT

PHRASES:

ADJACENT COMPONENTS	ONE INTERVENING WORD
COALI GOVERN, GOVERN FORM, FORM ITALY, ITALY LEFT-W, LEFT-W SOCIAL, SOCIAL REPUBLIC, ~~REPUBLIC SOCIAL~~*, SOCIAL DEMOCRAT, DEMOCRAT CHRIST, ~~CHRIST DEMOCRAT~~*	COALI FORM, GOVERN ITALY, FORM LEFT-W, ITALY SOCIAL, LEFT-W REPUBLIC, ~~SOCIAL SOCIAL~~,+ REPUBLIC DEMOCRAT, SOCIAL CHRIST, ~~DEMOCRAT DEMOCRAT~~+

* Duplicate Phrases Eliminated

\+ Identical Components Eliminated
Word Order Immaterial

Sample Phrase Formation Process

Fig. 8

Type of Terms	Fraction of Terms Covered (Fraction of Term Assignments)	Number of Terms	Document Frequency of Terms	Number of Good Discriminators From Top 50	Number of Good Discriminators From Top 100
Low Frequency	60%	CRAN 1668	1-3	0	0
		MED 3238	1-2	5	16
		TIME 8916	1-2	1	11
	70%	CRAN 1899	1-4	0	4
		MED 3600	1-3	8	25
		TIME 9944	1-3	2	13
	80%	CRAN 2153	1-9	5	14
		MED 3845	1-4	8	29
		TIME 11344	1-6	12	36
High Frequency	3.5% (42%)	CRAN 91	54-214	5	11
	(36%)	MED 163	20-138	5	6
	(45%)	TIME 492	36-271	8	13
	4.0% (45%)	CRAN 105	48-214	7	15
	(39%)	MED 192	18-138	10	14
	(47%)	TIME 555	33-271	8	13
	4.5% (47%)	CRAN 115	44-214	9	17
	(40%)	MED 206	17-138	10	14
	(51%)	TIME 660	29-271	8	13

Number of Good Discriminators for Various Deletion
Percentage of Low and High Frequency Terms
Table 1

	Minimum Document Frequency needed for High-Frequency Component	Average Document Frequency	
		Single Terms Entering Phrase Process	Phrases
CRANFIELD	(45)	106	39
MEDLARS	(22)	40	7
TIME	(49)	101	38

Average Document Frequency for Phrases

Table 2(a)

	Maximum Document Frequency needed for Thesaurus Class to Insure Inclusion	Average Document Frequency	
		Single Terms Entering Thesaurus Process	Thesaurus Classes
CRANFIELD	(60)	24	32
MEDLARS	(40)	10	16
TIME	(60)	17	31

Average Document Frequency for Thesaurus Classes

Table 2(b)

	CRAN 424			MED 450			TIME 425		
	Standard Term Frequency	Phrase Assignment	Advan-tage	Standard Term Frequency	Phrase Assignment	Advan-tage	Standard Term Frequency	Phrase Assignment	Advan-tage
.1	.6844	.8793	+28%	.7891	.8811	+12%	.7496	.8488	+13%
.2	.5303	.7344	+38%	.6750	.8149	+21%	.7071	.8419	+19%
.3	.4689	.6013	+28%	.5481	.6992	+28%	.6710	.7998	+19%
.4	.3482	.5205	+49%	.4807	.6481	+35%	.6452	.7729	+20%
.5	.3134	.4150	+32%	.4384	.5930	+35%	.6351	.7025	+11%
.6	.2556	.3623	+42%	.3721	.5450	+46%	.5866	.6800	+16%
.7	.1989	.3017	+52%	.3357	.4867	+45%	.5413	.6331	+17%
.8	.1631	.1953	+20%	.2195	.3263	+49%	.5004	.5895	+18%
.9	.1265	.1463	+16%	.1768	.2767	+56%	.3865	.4618	+19%
1.0	.1176	.1314	+12%	.1230	.1969	+60%	.3721	.4529	+22%
		Average	+32%		Average	+39%		Average	+17%

Average Precision Values at Ten Recall Points

(Phrase Process vs. Standard)

Table 3

	CRAN 424			MED 450			TIME 425		
	Standard Term Frequency	Thesaurus Plus Phrases	Advan-tage	Standard Term Frequency	Thesaurus Plus Phrases	Advan-tage	Standard Term Frequency	Thesaurus Plus Phrases	Advan-tage
.1	.6844	.8745	27.8%	.7891	.8919	13.0%	.7496	.8339	11.2%
.2	.5303	.7908	49.1%	.6750	.8331	23.4%	.7071	.8138	15.0%
.3	.4689	.6387	36.2%	.5481	.7057	28.8%	.6710	.7812	16.4%
.4	.3482	.5401	55.1%	.4807	.6443	34.0%	.6452	.7681	19.0%
.5	.3134	.4516	44.1%	.4384	.6099	39.1%	.6351	.7006	10.3%
.6	.2556	.3718	45.5%	.3721	.5548	49.1%	.5866	.6882	17.3%
.7	.1989	.2979	49.8%	.3357	.5179	54.3%	.5413	.6389	18.0%
.8	.1631	.2019	23.8%	.2195	.3949	79.9%	.5004	.5915	18.2%
.9	.1265	.1556	23.0%	.1768	.3505	98.2%	.3865	.4842	25.3%
1.0	.1176	.1375	16.9%	.1230	.2484	101.9%	.3721	.4790	28.7%
		Average	+37%		Average	+52%		Average	+18%
		Average (Phrases)	+32%		Average (Phrases)	+39%		Average (Phrases)	+17%
			+ 5%			+ 13%			+ 1%

Average Precision Values at Ten Recall Points

(Thesaurus and Phrases vs. Standard)

Table 4

CRAN 424	
Automatic Phrases vs. Standard Term Frequency +32%	Automa vs Fr
Automatic Phrases Plus Thesaurus vs. Standard Run +37%	Automa Pl vs
Best Precision Low Recall 0.89 Medium Recall 0.43 High Recall 0.13	Best L Medi Hi

Summary of Recall-Precis

MED 450	TIME 425
tic Phrases . Standard Term equency +39%	Automatic Phrases vs. Standard Term Frequency +17%
tic Phrases us Thesaurus . Standard Run +52%	Automatic Phrases Plus Thesaurus vs. Standard Run +18%
recision w Recall 0.88 m Recall 0.61 h Recall 0.23	Best Precision Low Recall 0.85 Medium Recall 0.70 High Recall 0.45

ion Evaluation (Three Collections)

Table 5