

DOCUMENT RESUME**ED 096 987****IR 001 168**

AUTHOR Salton, G.; And Others
TITLE A Theory of Term Importance in Automatic Text Analysis.
INSTITUTION Cornell Univ., Ithaca, N.Y. Dept. of Computer Science.
PUB DATE 74
NOTE 18p.; This document may not reproduce clearly due to small size of type

EDRS PRICE MF-\$0.75 HC-\$1.50 PLUS POSTAGE
DESCRIPTORS *Automatic Indexing; *Automation; *Content Analysis; Information Retrieval; Information Scientists; Models; Problem Solving; Thesauri
IDENTIFIERS Discrimination Value Analysis; Space Density; Vectors

ABSTRACT

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)

U.S. DEPARTMENT OF HEALTH,
EDUCATION & WELFARE
NATIONAL INSTITUTE OF
EDUCATION
THIS DOCUMENT HAS BEEN REPRO-
DUCED EXACTLY AS RECEIVED FROM
THE PERSON OR ORGANIZATION ORIGIN-
ATING IT POINTS OF VIEW OR OPINIONS
STATED DO NOT NECESSARILY REPRESENT
THE OFFICIAL NATIONAL INSTITUTE OF
EDUCATION POSITION OR POLICY

BEST COPY AVAILABLE

A Theory of Term Importance in Automatic

Text Analysis

G. Salton[†], C.S. Yang[‡], and C.T. Yu[†]

The theory is validated by citing experimental results.

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:

- the role and importance of single words;
- the role of juxtaposed words (phrases);
- 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

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

where w_{ij} represents the weight of term j in the vector corresponding to the i th 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.

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. [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

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

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 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.

BEST COPY AVAILABLE

value of term k may then be defined as

$$DV_k = Q_k - Q. \quad (3)$$

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 (Tranfield collection, 424 documents comprising 2551 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 ,

$$w_{kj} = f_{kj} \cdot DV_k, \quad (4)$$

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 $d_k = 1$. 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.

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 300 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.

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 $n/20$ 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.

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 8915 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

BEST COPY AVAILABLE

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 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:

BEST COPY AVAILABLE

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. 9 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 in 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

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]

$$w_{ipq} = \frac{w_{ip} + w_{iq}}{2} \tag{5}$$

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:

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

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 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)):

$$w_{ipq} = \left[\log n - \frac{\log d_p + \log d_q}{2} \right] \cdot \left[\frac{w_{ip} + w_{iq}}{2} \right] \tag{7}$$

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 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 occurred 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. [5,5,7,8] 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 that 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_{ij}) for phrase 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.

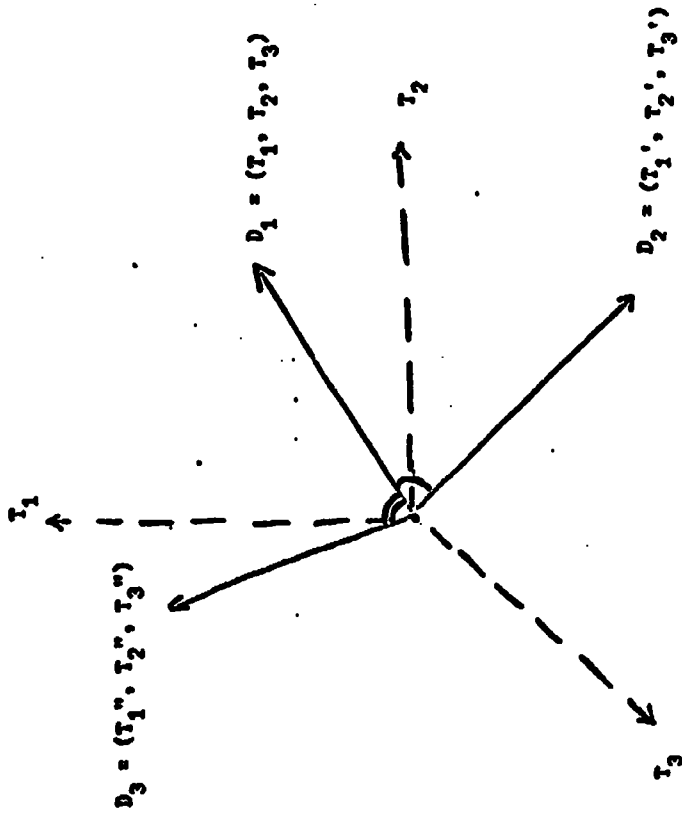
Table 4 contains output similar to that already shown in Table 3. However the data in Table 4 apply to an indexing system using both left-to-right (thesaurus) and right-to-left (phrase) transformations. It is seen from Table 4 that the thesaurus transformation adds an additional average improvement of 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, 0.40, 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 now 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. 23, 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 K.E. Lesk, Computer Evaluation of Indexing and Text Processing, *Journal of the ACM*, Vol. 15, No. 1, January 1968, p. 9-36.
- [4] K. Sparck Jones, A Statistical Interpretation of Term Specificity and its Application to Retrieval, *Journal of Documentation*, Vol. 23, No. 1, March 1972, p. 11-20.
- [5] K. Sparck Jones, *Automatic Keyword Classifications*, Butterworths, London, 1971.
- [6] C.C. Gottlieb 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 Automatic Proc. IFF Conference, Stockholm, August 1974.

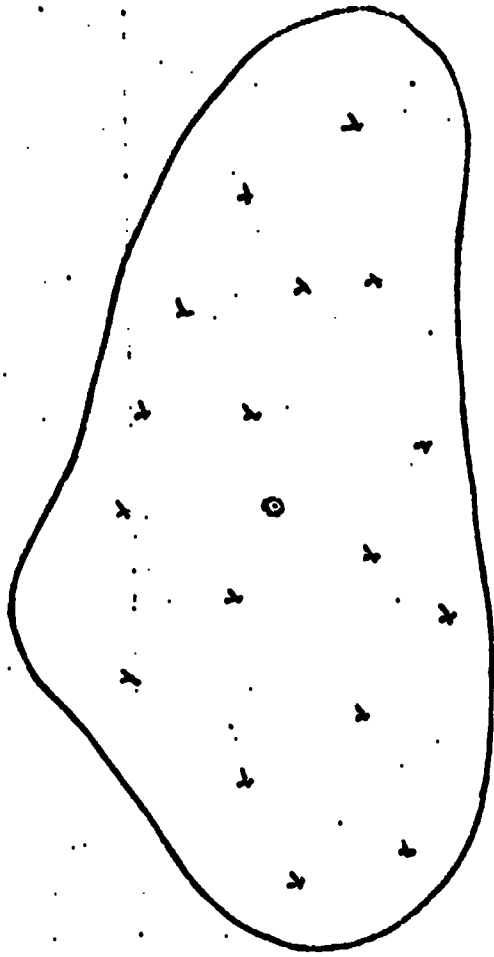
BEST COPY AVAILABLE



Vector Representation of Document Space

Fig. 1

BEST COPY AVAILABLE



○ Centroid of Space
x Individual Document

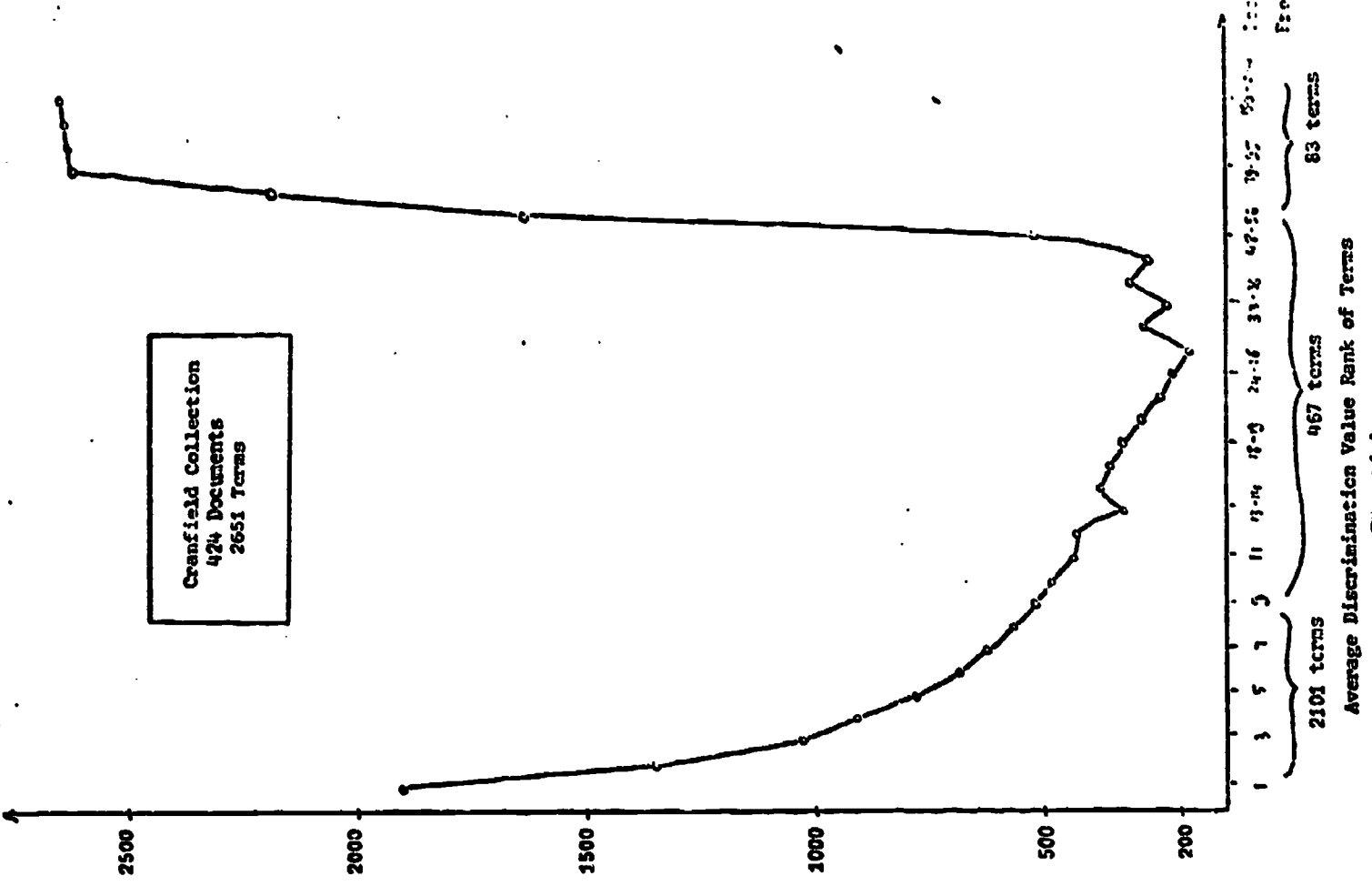
Multidimensional Document Space

Fig. 2

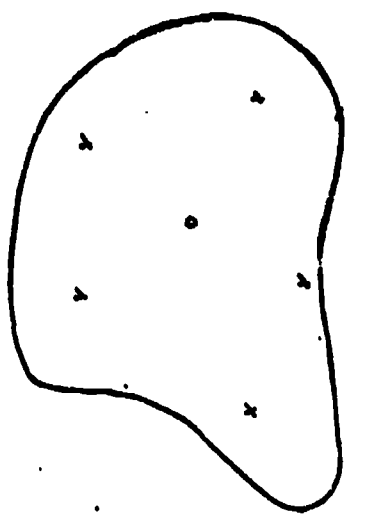
BEST COPY AVAILABLE

22
Discrimination Rank
of Average Term

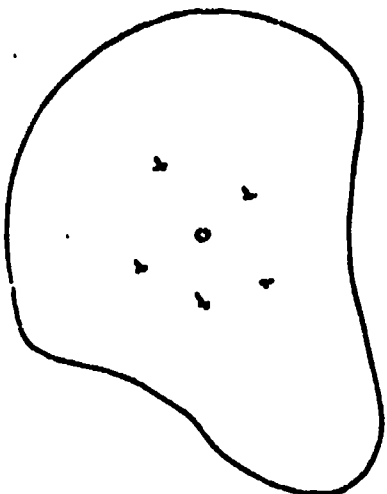
Cranfield Collection
424 Documents
2651 Terms



21



After Assignment
of Term



Before Assignment
of Term

- x Document
- o Main Centroid

Operation of
Good Discriminating Term

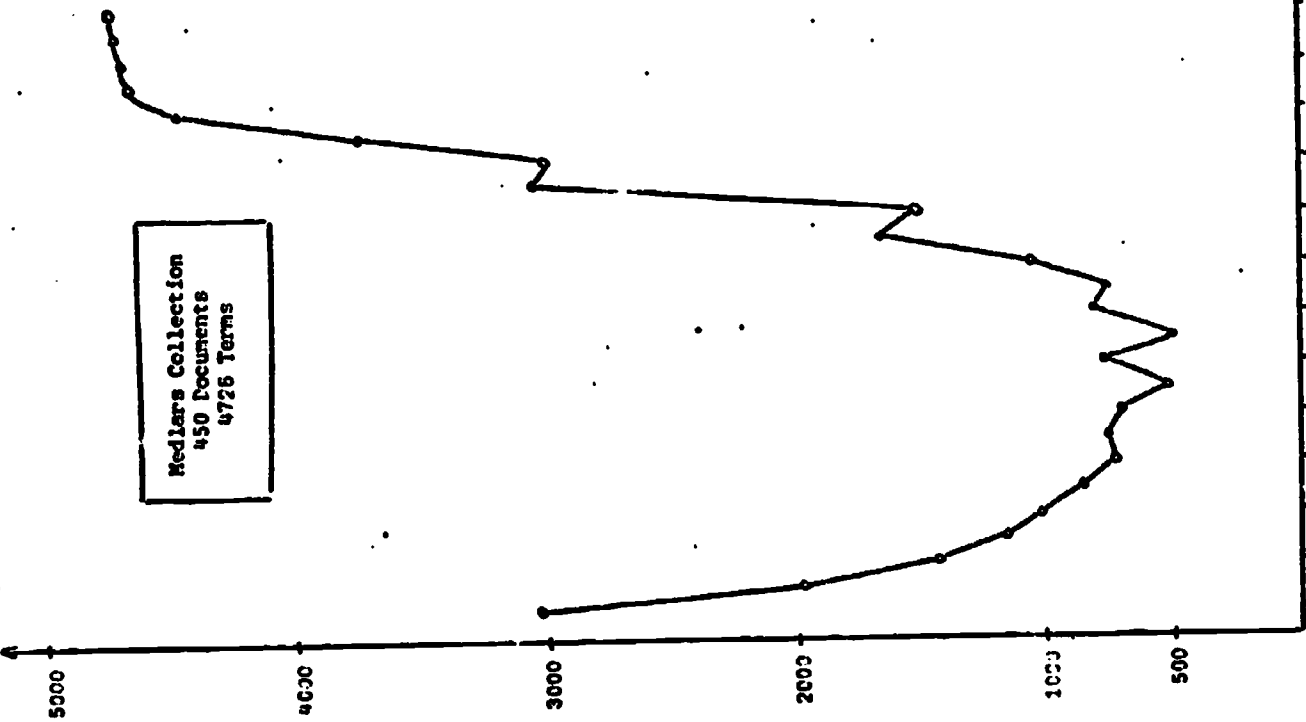
Fig. 3

BEST COPY AVAILABLE

Discrimination Rank
Average Term

23

Medlars Collection
450 Documents
4726 Terms



Document
Frequency

1 3 5 7 9 11 13 15 17 19 21 23 25 27 29 31 33 35 37 39 41 43 45 47 49 51 53 55 57 59 61 63 65 67 69 71 73 75 77 79 81 83 85 87 89 91 93 95 97 99 101 103 105 107 109 111 113 115 117 119 121 123 125 127 129 131 133 135 137 139 141 143 145 147 149 151 153 155 157 159 161 163 165 167 169 171 173 175 177 179 181 183 185 187 189 191 193 195 197 199 201 203 205 207 209 211 213 215 217 219 221 223 225 227 229 231 233 235 237 239 241 243 245 247 249 251 253 255 257 259 261 263 265 267 269 271 273 275 277 279 281 283 285 287 289 291 293 295 297 299 301 303 305 307 309 311 313 315 317 319 321 323 325 327 329 331 333 335 337 339 341 343 345 347 349 351 353 355 357 359 361 363 365 367 369 371 373 375 377 379 381 383 385 387 389 391 393 395 397 399 401 403 405 407 409 411 413 415 417 419 421 423 425 427 429 431 433 435 437 439 441 443 445 447 449 451 453 455 457 459 461 463 465 467 469 471 473 475 477 479 481 483 485 487 489 491 493 495 497 499 501 503 505 507 509 511 513 515 517 519 521 523 525 527 529 531 533 535 537 539 541 543 545 547 549 551 553 555 557 559 561 563 565 567 569 571 573 575 577 579 581 583 585 587 589 591 593 595 597 599 601 603 605 607 609 611 613 615 617 619 621 623 625 627 629 631 633 635 637 639 641 643 645 647 649 651 653 655 657 659 661 663 665 667 669 671 673 675 677 679 681 683 685 687 689 691 693 695 697 699 701 703 705 707 709 711 713 715 717 719 721 723 725 727 729 731 733 735 737 739 741 743 745 747 749 751 753 755 757 759 761 763 765 767 769 771 773 775 777 779 781 783 785 787 789 791 793 795 797 799 801 803 805 807 809 811 813 815 817 819 821 823 825 827 829 831 833 835 837 839 841 843 845 847 849 851 853 855 857 859 861 863 865 867 869 871 873 875 877 879 881 883 885 887 889 891 893 895 897 899 901 903 905 907 909 911 913 915 917 919 921 923 925 927 929 931 933 935 937 939 941 943 945 947 949 951 953 955 957 959 961 963 965 967 969 971 973 975 977 979 981 983 985 987 989 991 993 995 997 999 1001 1003 1005 1007 1009 1011 1013 1015 1017 1019 1021 1023 1025 1027 1029 1031 1033 1035 1037 1039 1041 1043 1045 1047 1049 1051 1053 1055 1057 1059 1061 1063 1065 1067 1069 1071 1073 1075 1077 1079 1081 1083 1085 1087 1089 1091 1093 1095 1097 1099 1101 1103 1105 1107 1109 1111 1113 1115 1117 1119 1121 1123 1125 1127 1129 1131 1133 1135 1137 1139 1141 1143 1145 1147 1149 1151 1153 1155 1157 1159 1161 1163 1165 1167 1169 1171 1173 1175 1177 1179 1181 1183 1185 1187 1189 1191 1193 1195 1197 1199 1201 1203 1205 1207 1209 1211 1213 1215 1217 1219 1221 1223 1225 1227 1229 1231 1233 1235 1237 1239 1241 1243 1245 1247 1249 1251 1253 1255 1257 1259 1261 1263 1265 1267 1269 1271 1273 1275 1277 1279 1281 1283 1285 1287 1289 1291 1293 1295 1297 1299 1301 1303 1305 1307 1309 1311 1313 1315 1317 1319 1321 1323 1325 1327 1329 1331 1333 1335 1337 1339 1341 1343 1345 1347 1349 1351 1353 1355 1357 1359 1361 1363 1365 1367 1369 1371 1373 1375 1377 1379 1381 1383 1385 1387 1389 1391 1393 1395 1397 1399 1401 1403 1405 1407 1409 1411 1413 1415 1417 1419 1421 1423 1425 1427 1429 1431 1433 1435 1437 1439 1441 1443 1445 1447 1449 1451 1453 1455 1457 1459 1461 1463 1465 1467 1469 1471 1473 1475 1477 1479 1481 1483 1485 1487 1489 1491 1493 1495 1497 1499 1501 1503 1505 1507 1509 1511 1513 1515 1517 1519 1521 1523 1525 1527 1529 1531 1533 1535 1537 1539 1541 1543 1545 1547 1549 1551 1553 1555 1557 1559 1561 1563 1565 1567 1569 1571 1573 1575 1577 1579 1581 1583 1585 1587 1589 1591 1593 1595 1597 1599 1601 1603 1605 1607 1609 1611 1613 1615 1617 1619 1621 1623 1625 1627 1629 1631 1633 1635 1637 1639 1641 1643 1645 1647 1649 1651 1653 1655 1657 1659 1661 1663 1665 1667 1669 1671 1673 1675 1677 1679 1681 1683 1685 1687 1689 1691 1693 1695 1697 1699 1701 1703 1705 1707 1709 1711 1713 1715 1717 1719 1721 1723 1725 1727 1729 1731 1733 1735 1737 1739 1741 1743 1745 1747 1749 1751 1753 1755 1757 1759 1761 1763 1765 1767 1769 1771 1773 1775 1777 1779 1781 1783 1785 1787 1789 1791 1793 1795 1797 1799 1801 1803 1805 1807 1809 1811 1813 1815 1817 1819 1821 1823 1825 1827 1829 1831 1833 1835 1837 1839 1841 1843 1845 1847 1849 1851 1853 1855 1857 1859 1861 1863 1865 1867 1869 1871 1873 1875 1877 1879 1881 1883 1885 1887 1889 1891 1893 1895 1897 1899 1901 1903 1905 1907 1909 1911 1913 1915 1917 1919 1921 1923 1925 1927 1929 1931 1933 1935 1937 1939 1941 1943 1945 1947 1949 1951 1953 1955 1957 1959 1961 1963 1965 1967 1969 1971 1973 1975 1977 1979 1981 1983 1985 1987 1989 1991 1993 1995 1997 1999 2001 2003 2005 2007 2009 2011 2013 2015 2017 2019 2021 2023 2025 2027 2029 2031 2033 2035 2037 2039 2041 2043 2045 2047 2049 2051 2053 2055 2057 2059 2061 2063 2065 2067 2069 2071 2073 2075 2077 2079 2081 2083 2085 2087 2089 2091 2093 2095 2097 2099 2101 2103 2105 2107 2109 2111 2113 2115 2117 2119 2121 2123 2125 2127 2129 2131 2133 2135 2137 2139 2141 2143 2145 2147 2149 2151 2153 2155 2157 2159 2161 2163 2165 2167 2169 2171 2173 2175 2177 2179 2181 2183 2185 2187 2189 2191 2193 2195 2197 2199 2201 2203 2205 2207 2209 2211 2213 2215 2217 2219 2221 2223 2225 2227 2229 2231 2233 2235 2237 2239 2241 2243 2245 2247 2249 2251 2253 2255 2257 2259 2261 2263 2265 2267 2269 2271 2273 2275 2277 2279 2281 2283 2285 2287 2289 2291 2293 2295 2297 2299 2301 2303 2305 2307 2309 2311 2313 2315 2317 2319 2321 2323 2325 2327 2329 2331 2333 2335 2337 2339 2341 2343 2345 2347 2349 2351 2353 2355 2357 2359 2361 2363 2365 2367 2369 2371 2373 2375 2377 2379 2381 2383 2385 2387 2389 2391 2393 2395 2397 2399 2401 2403 2405 2407 2409 2411 2413 2415 2417 2419 2421 2423 2425 2427 2429 2431 2433 2435 2437 2439 2441 2443 2445 2447 2449 2451 2453 2455 2457 2459 2461 2463 2465 2467 2469 2471 2473 2475 2477 2479 2481 2483 2485 2487 2489 2491 2493 2495 2497 2499 2501 2503 2505 2507 2509 2511 2513 2515 2517 2519 2521 2523 2525 2527 2529 2531 2533 2535 2537 2539 2541 2543 2545 2547 2549 2551 2553 2555 2557 2559 2561 2563 2565 2567 2569 2571 2573 2575 2577 2579 2581 2583 2585 2587 2589 2591 2593 2595 2597 2599 2601 2603 2605 2607 2609 2611 2613 2615 2617 2619 2621 2623 2625 2627 2629 2631 2633 2635 2637 2639 2641 2643 2645 2647 2649 2651 2653 2655 2657 2659 2661 2663 2665 2667 2669 2671 2673 2675 2677 2679 2681 2683 2685 2687 2689 2691 2693 2695 2697 2699 2701 2703 2705 2707 2709 2711 2713 2715 2717 2719 2721 2723 2725 2727 2729 2731 2733 2735 2737 2739 2741 2743 2745 2747 2749 2751 2753 2755 2757 2759 2761 2763 2765 2767 2769 2771 2773 2775 2777 2779 2781 2783 2785 2787 2789 2791 2793 2795 2797 2799 2801 2803 2805 2807 2809 2811 2813 2815 2817 2819 2821 2823 2825 2827 2829 2831 2833 2835 2837 2839 2841 2843 2845 2847 2849 2851 2853 2855 2857 2859 2861 2863 2865 2867 2869 2871 2873 2875 2877 2879 2881 2883 2885 2887 2889 2891 2893 2895 2897 2899 2901 2903 2905 2907 2909 2911 2913 2915 2917 2919 2921 2923 2925 2927 2929 2931 2933 2935 2937 2939 2941 2943 2945 2947 2949 2951 2953 2955 2957 2959 2961 2963 2965 2967 2969 2971 2973 2975 2977 2979 2981 2983 2985 2987 2989 2991 2993 2995 2997 2999 3001 3003 3005 3007 3009 3011 3013 3015 3017 3019 3021 3023 3025 3027 3029 3031 3033 3035 3037 3039 3041 3043 3045 3047 3049 3051 3053 3055 3057 3059 3061 3063 3065 3067 3069 3071 3073 3075 3077 3079 3081 3083 3085 3087 3089 3091 3093 3095 3097 3099 3101 3103 3105 3107 3109 3111 3113 3115 3117 3119 3121 3123 3125 3127 3129 3131 3133 3135 3137 3139 3141 3143 3145 3147 3149 3151 3153 3155 3157 3159 3161 3163 3165 3167 3169 3171 3173 3175 3177 3179 3181 3183 3185 3187 3189 3191 3193 3195 3197 3199 3201 3203 3205 3207 3209 3211 3213 3215 3217 3219 3221 3223 3225 3227 3229 3231 3233 3235 3237 3239 3241 3243 3245 3247 3249 3251 3253 3255 3257 3259 3261 3263 3265 3267 3269 3271 3273 3275 3277 3279 3281 3283 3285 3287 3289 3291 3293 3295 3297 3299 3301 3303 3305 3307 3309 3311 3313 3315 3317 3319 3321 3323 3325 3327 3329 3331 3333 3335 3337 3339 3341 3343 3345 3347 3349 3351 3353 3355 3357 3359 3361 3363 3365 3367 3369 3371 3373 3375 3377 3379 3381 3383 3385 3387 3389 3391 3393 3395 3397 3399 3401 3403 3405 3407 3409 3411 3413 3415 3417 3419 3421 3423 3425 3427 3429 3431 3433 3435 3437 3439 3441 3443 3445 3447 3449 3451 3453 3455 3457 3459 3461 3463 3465 3467 3469 3471 3473 3475 3477 3479 3481 3483 3485 3487 3489 3491 3493 3495 3497 3499 3501 3503 3505 3507 3509 3511 3513 3515 3517 3519 3521 3523 3525 3527 3529 3531 3533 3535 3537 3539 3541 3543 3545 3547 3549 3551 3553 3555 3557 3559 3561 3563 3565 3567 3569 3571 3573 3575 3577 3579 3581 3583 3585 3587 3589 3591 3593 3595 3597 3599 3601 3603 3605 3607 3609 3611 3613 3615 3617 3619 3621 3623 3625 3627 3629 3631 3633 3635 3637 3639 3641 3643 3645 3647 3649 3651 3653 3655 3657 3659 3661 3663 3665 3667 3669 3671 3673 3675 3677 3679 3681 3683 3685 3687 3689 3691 3693 3695 3697 3699 3701 3703 3705 3707 3709 3711 3713 3715 3717 3719 3721 3723 3725 3727 3729 3731 3733 3735 3737 3739 3741 3743 3745 3747 3749 3751 3753 3755 3757 3759 3761 3763 3765 3767 3769 3771 3773 3775 3777 3779 3781 3783 3785 3787 3789 3791 3793 3795 3797 3799 3801 3803 3805 3807 3809 3811 3813 3815 3817 3819 3821 3823 3825 3827 3829 3831 3833 3835 3837 3839 3841 3843 3845 3847 3849 3851 3853 3855 3857 3859 3861 3863 3865 3867 3869 3871 3873 3875 3877 3879 3881 3883 3885 3887 3889 3891 3893 3895 3897 3899 3901 3903 3905 3907 3909 3911 3913 3915 3917 3919 3921 3923 3925 3927 3929 3931 3933 3935 3937 3939 3941 3943 3945 3947 3949 3951 3953 3955 3957 3959 3961 3963 3965 3967 3969 3971 3973 3975 3977 3979 3981 3983 3985 3987 3989 3991 3993 3995 3997 3999 4001 4003 4005 4007 4009 4011 4013 4015 4017 4019 4021 4023 4025 4027 4029 4031 4033 4035 4037 4039 4041 4043 4045 4047 4049 4051 4053 4055 4057 4059 4061 4063 4065 4067 4069 4071 4073 4075 4077 4079 4081 4083 4085 4087 4089 4091 4093 4095 4097 4099 4101 4103 4105 4107 4109 4111 4113 4115 4117 4119 4121 4123 4125 4127 4129 4131 4133 4135 4137 4139 4141 4143 4145 4147 4149 4151 4153 4155 4157 4159 4161 4163 4165 4167 4169 4171 4173 4175 4177 4179 4181 4183 4185 4187 4189 4191 4193 4195 4197 4199 4201 4203 4205 4207 4209 4211 4213 4215 4217 4219 4221 4223 4225 4227 4229 4231 4233 4235 4237 4239 4241 4243 4245 4247 4249 4251 4253 4255 4257 4259 4261 4263 4265 4267 4269 4271 4273 4275 4277 4279 4281 4283 4285 4287 4289 4291 4293 4295 4297 4299 4301 4303 4305 4307 4309 4311 4313 4315 4317 4319 4321 4323 4325 4327 4329 4331 4333 4335 4337 4339 4341 4343 4345 4347 4349 4351 4353 4355 4357 4359 4361 4363 4365 4367 4369 4371 4373 4375 4377 4379 4381 4383 4385 4387 4389 4391 4393 4395 4397 4399 4401 4403 4405 4407 4409 4411 4413 4415 4417 4419 4421 4423 4425 4427 4429 4431 4433 4435 4437 4439 4441 4443 4445 4447 4449 4451 4453 4455 4457 4459 4461 4463 4465 4467 4469 4471 4473 4475 4477 4479 4481 4483 4485 4487 4489 4491 4493 4495 4497 4499 4501 4503 4505 4507 4509 4511 4513 4515 4517 4519 4521 4523 4525 4527 4529 4531 4533 4535 4537 4539 4541 4543 4545 4547 4549 4551 4553 4555 4557 4559 4561 4563 4565 4567 4569 4571 4573 4575 4577 4579 4581 4583 4585 4587 4589 4591 4593 4595 4597 4599 4601 4603 4605 4607 4609 4611 4613 4615 4617 4619 4621 4623 4625 4627 4629 4631 4633 4635 4637 4639 4641 4643 4645 4647 4649 4651 4653 4655 4657 4659 4661 4663 4665 4667 4669 4671 4673 4675 4677 4679 4681 4683 4685 4687 4689 4691 4693 4695 4697 4699 4701 4703 4705 4707 4709 4711 4713 4715 4717 4719 4721 4723 4725 4727 4729 4731 4733 4735 4737 4739 4741 4743 4745 4747 4749 4751 4753 4755 4757 4759 4761 4763 4765 4767 4769 4771 4773 4775 4777 4779 4781 4783 4785 4787 4789 4791 4793 4795 4797 4799 4801 4803 4805 4807 4809 4811 4813 4815 4817 4819 4821 4823 4825 4827 4829 4831 4833 4835 4837 4839 4841 4843 4845 4847 4849 48

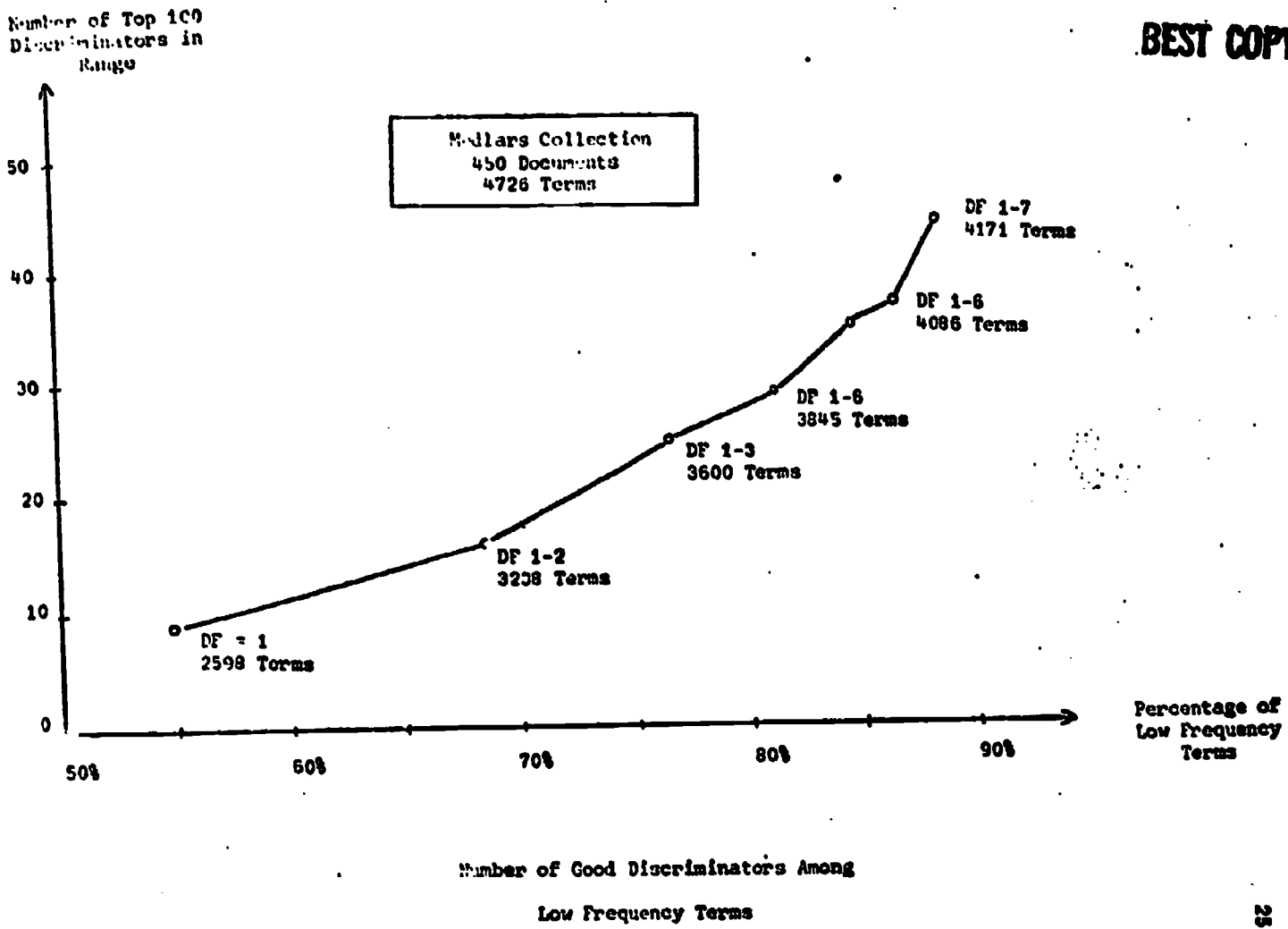


Fig. 5(a)

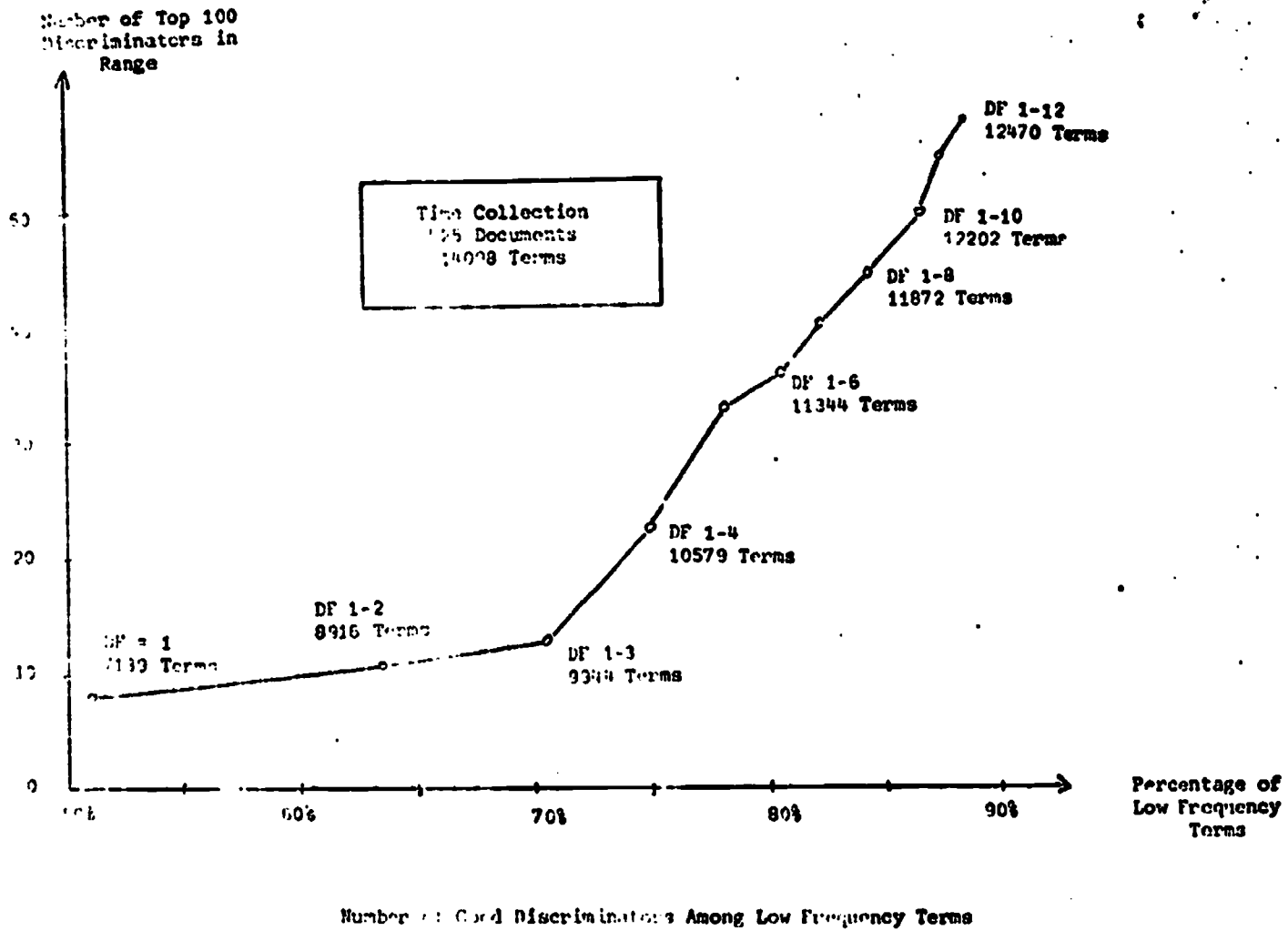


Fig. 5(b)

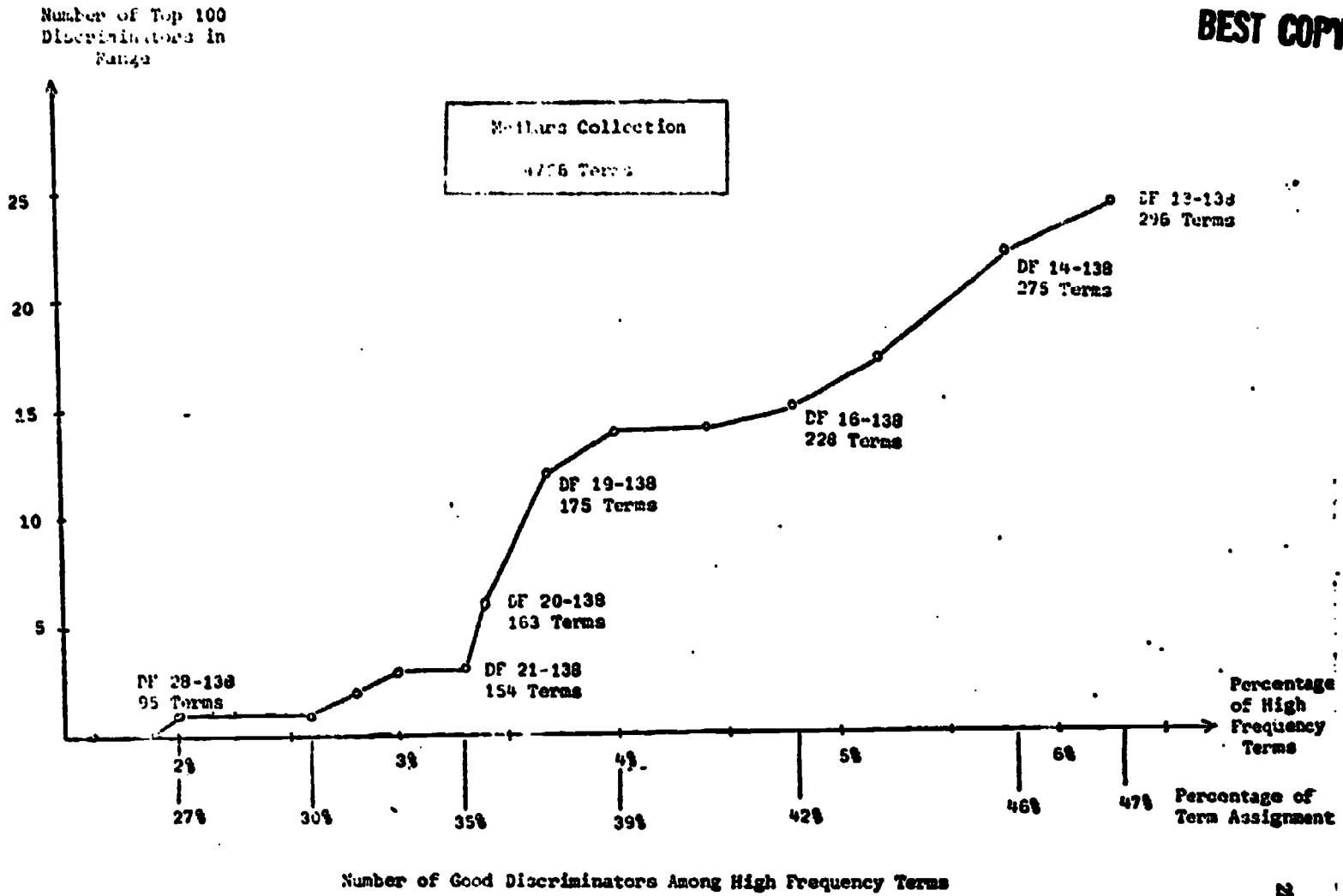


Fig. 6(a)

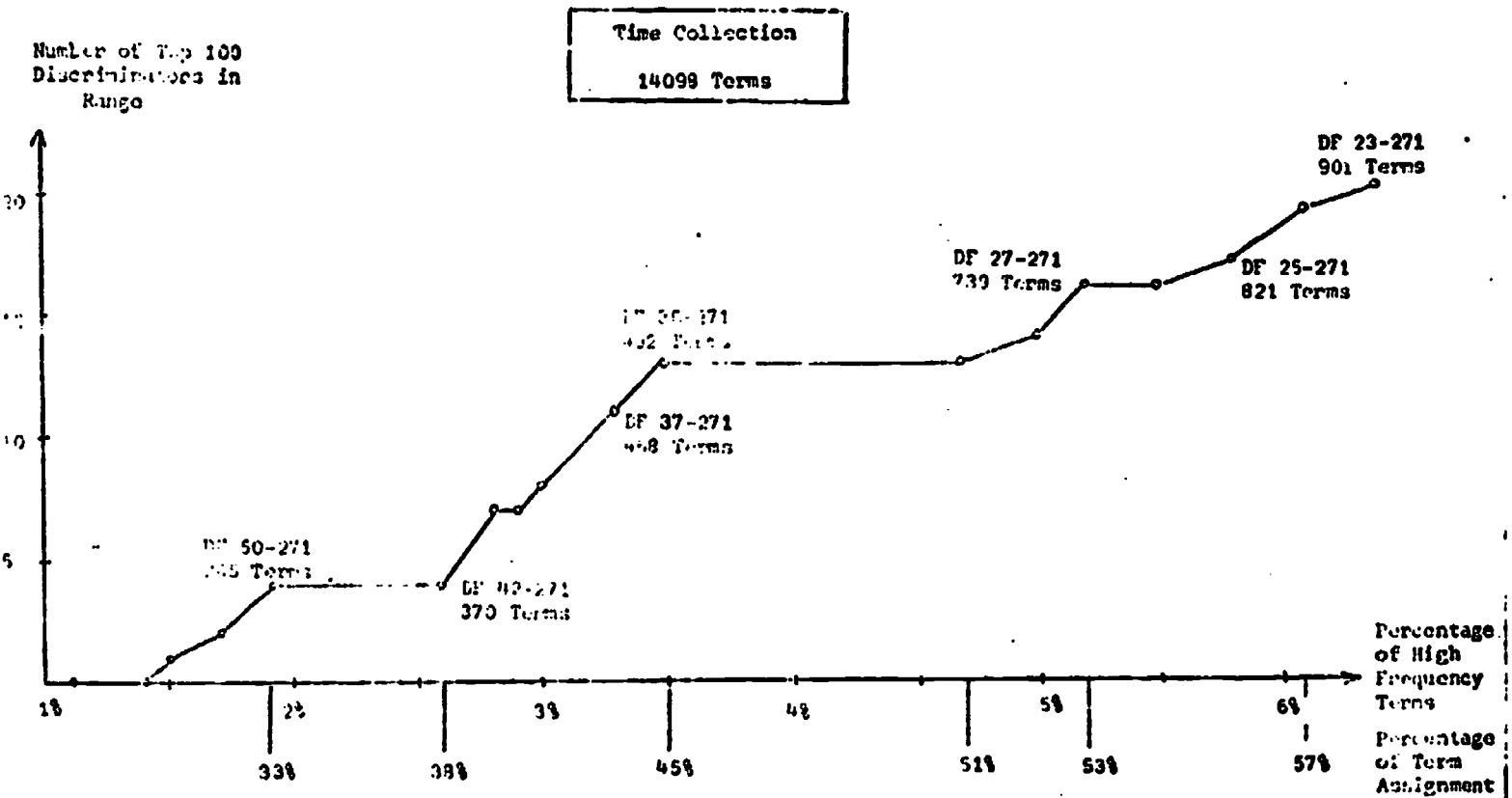
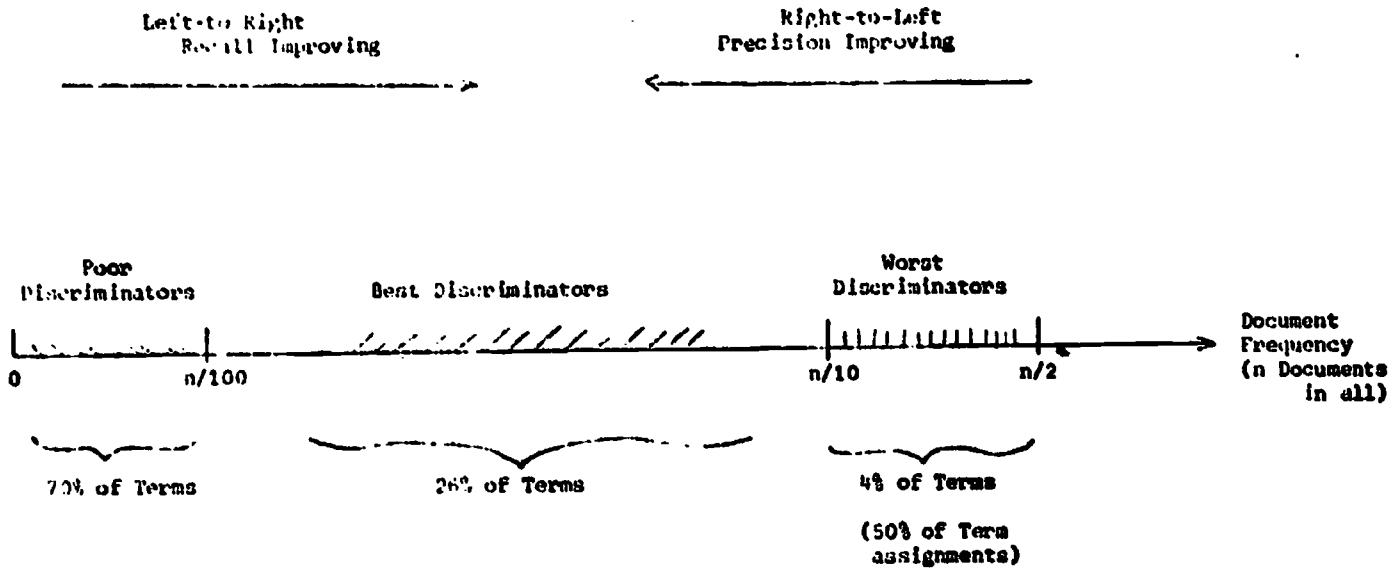


Fig. 6(b)



Summarization of Discrimination Value of Terms in Frequency Ranges

Fig. 7

28

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:

COALIT GOVERN FORM ITALY LEFT-W SOCIAL REPUBLICAN
SOCIAL DEMOCRAT CHRIST DEMOCRAT

PHRASES:

<u>ADJACENT COMPONENTS</u>	<u>ONE INTERPRETING WORD</u>
COALIT GOVERN, GOVERN FORM, FORM ITALY, ITALY LEFT-W, LEFT-W SOCIAL, SOCIAL REPUBLICAN, SOCIAL DEMOCRAT, CHRIST DEMOCRAT, CHRISTIAN DEMOCRAT	COALIT FORM, GOVERN ITALY, FORM LEFT-W, ITALY SOCIAL, LEFT-W REPUBLICAN, SOCIAL DEMOCRAT, CHRISTIAN DEMOCRAT

- * Duplicate Phrases Eliminated
- * Identical Components Identified and Deleted

Sample Phrase Formation Process

Fig. 9

30

BEST COPY AVAILABLE

	Minimum Document Frequency needed for High-Frequency Component	Average Document Frequency	
		Single Terms Entering Phrase Process	Phrases
CRANFIELD	(45)	106	33
MEDLARS	(22)	40	7
TIME	(49)	101	38

Average Document Frequency for Phrases
Table 2(a)

	Maximum Document Frequency needed for Thesaurus Class to Include 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)

Type of Term	Fraction of Terms Covered (Fraction of Term Assignments)	Number of Terms	Document Frequency of Terms	Number of Good Discriminators	
				From Top 50	From Top 100
Low Frequency	60%	CRAN 1666	1-3	0	0
		MED 3236	1-2	5	16
		TIME 8916	1-2	1	11
	70%	CRAN 1999	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	2-6	12	36	
High Frequency	3.5% (36%)	CRAN 91	54-214	5	11
		MED 163	20-138	5	6
		TIME 492	36-271	8	13
	4.5% (45%)	CRAN 105	48-214	7	15
		MED 192	18-138	10	14
		TIME 555	33-271	8	13
4.0% (40%)	CRAN 115	44-214	9	17	
	MED 206	17-138	10	14	
	TIME 660	29-271	8	13	

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

	SPAN 424			MED 450			TIME 425		
	Standard Term Frequency	Phrase Assignment	Advantage	Standard Term Frequency	Phrase Assignment	Advantage	Standard Term Frequency	Phrase Assignment	Advantage
.1	.6844	.8793	+28%	.7891	.8911	+12%	.7496	.8408	+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	+15%	.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

33

BEST COPY AVAILABLE

	SPAN 424			MED 450			TIME 425		
	Standard Term Frequency	Thesaurus Plus Phrases	Advantage	Standard Term Frequency	Thesaurus Plus Phrases	Advantage	Standard Term Frequency	Thesaurus Plus Phrases	Advantage
.1	.6844	.8745	27.3%	.7891	.8919	13.0%	.7496	.8339	11.2%
.2	.5303	.7108	33.1%	.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	45.3%	.3721	.5548	49.1%	.5866	.6882	17.3%
.7	.1989	.2779	49.9%	.3357	.5179	54.3%	.5413	.6389	18.0%
.8	.1631	.2019	23.9%	.2195	.3949	79.7%	.5004	.5915	18.2%
.9	.1265	.1556	21.0%	.1768	.3505	98.2%	.3865	.4842	25.3%
1.0	.1176	.1375	16.5%	.1230	.2484	101.9%	.3721	.4790	28.7%
	Average +17%			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

34

CRAN 424	MED 450	TIME 425
<p>Automatic Phrases vs. Standard Term Frequency <u>+32%</u></p>	<p>Automatic Phrases vs. Standard Term Frequency <u>+39%</u></p>	<p>Automatic Phrases vs. Standard Term Frequency <u>+17%</u></p>
<p>Automatic Phrases Plus Thesaurus vs. Standard Run <u>+37%</u></p>	<p>Automatic Phrases Plus Thesaurus vs. Standard Run <u>+52%</u></p>	<p>Automatic Phrases Plus Thesaurus vs. Standard Run <u>+18%</u></p>
<p>Best Precision Low Recall 0.89 Medium Recall 0.43 High Recall 0.13</p>	<p>Best Precision Low Recall 0.88 Medium Recall 0.61 High Recall 0.23</p>	<p>Best Precision Low Recall 0.85 Medium Recall 0.70 High Recall 0.45</p>

Summary of Recall-Precision Evaluation (Three Collections)

Table 5