ON THE SPECIFICATION OF TERM VALUES IN AUTOMATIC INDEXING

G. Salton

C.S. Yang

TR 73-173

June 1973

Revised 7/2/73

Department of Computer Science
Cornell University
Ithaca, New York 14850

On the Specification of Term Values in Automatic Indexing

G. Salton and C.S. Yang[+]

Abstract

The existing practice in automatic indexing is reviewed, and it is shown that the standard theories for the specification of term values (or weights) are not adequate. New techniques are introduced for the assignment of weights to index terms, based on the characteristics of individual document collections. The effectiveness of some of the proposed methods is evaluated.

1. Current Indexing Practice

Two fundamental notions in the theory of automatic indexing are known respectively as indexing exhaustivity and term specificity. Indexing exhaustivity refers to the accuracy and depth with which the various topic areas germane to a given document are reflected in the set of index terms assigned to the document, whereas term specificity is a function of the exactness with which a term characterizes a given subject. In general, increasing exhaustivity implies a better recall performance, while increasing term specificity means better precision. In particular, the more exhaustive the indexing, that is, the more thorough the coverage of the various subject areas, the more likely it is that relevant items are actually retrieved in response

+ Department of Computer Science, Cornell University, Ithaca, New York 14850

to user queries, thus achieving high recall; similarly, the greater the term specificity, that is, the more precise the definition of each term, the less likely it is that extraneous nonrelevant items are also retrieved, thus achieving high precision. In a given user and collection context, one must then look for an optimum level of specificity in the vocabulary, and an optimum level of exhaustivity in the indexing to cover the recall and/or precision performance desired by the user population.

In an actual operating environment, one may conjecture that indexing exhaustivity has something to do with the number of index terms assigned to a given document, particularly the number of higher frequency terms — those largely responsible for the recall performance. Term specificity, on the other hand may be assumed to be related to the number of documents to which a given term is assigned in a given collection, the idea being that the smaller the document frequency, that is, the more concentrated the assignment of a term to only a few documents in a collection, the more likely it is that a given term is reasonably specific. [1]

The introduction of relationships between the indexing exhaustivity and specificity on the one hand, and the frequency characteristics of the index terms on the other, has led to certain indexing theories which have been used widely in practice. Before reviewing the main theories, it is convenient to distinguish two different frequency measures. The term frequency f_i^k is the frequency of occurrence of term i in document k. The total frequency of occurrence, F_i, of term i is then defined simply as the sum of the individual term frequencies across the N documents of a collection, that is,

$$F_i = \sum_{i=1}^{N} f_i^k \quad . \qquad (1)$$

A somewhat different measure is the document frequency d_i of term i which measures the number of documents to which term i is assigned. In an indexing system in which no weights are assigned to the terms, that is, where f_i^k is equal to 1 for all k and all i whenever term i appears in document k, and f_i^k is zero otherwise, the document frequency d_i then equals the total frequency F_i for all i.

Based on the concepts of term and document frequencies, a large variety of indexing methods can be implemented using completely objective criteria which depend only on the occurrence characteristics of terms in documents. The first and best known of these is due to Luhn, and assumes that the value, or weight, of a term, assigned to a document is simply proportional to the term frequency (TF); that is, the more often a term occurs in the text of a document, the higher its weight. [2] The Luhn theory reflects the fact that high frequency terms are often essential for the specification of document content and for the retrieval of relevant information.

In many environments, the standard term frequency weights do, in fact, enhance the retrieval performance, particularly at the high recall end of the performance curve, as shown in the example of Fig. 1 for a collection of 425 documents in world affairs taken from issues

of Time magazine published in 1963, and processed against 24 user queries.* It may be seen that the performance using binary weights (where all term weights equal 1 when a term is present in a document, and 0 otherwise) is inferior by as much as 25 percent in precision at high recall to that obtained with regular term frequencies.

Unfortunately, the term frequency weighting does not always perform as expected. In particular, when few high frequency index terms are present in a given collection, or when the high frequency terms are evenly distributed across the documents — for example, when a given term occurs k times in each of the documents — the upweighting of the high frequency terms will be of no avail. The output of Fig. 2, comparing binary with term frequency weights for a collection of 450 document abstracts in biomedicine, called MED, is a case in point. It may be seen in Fig. 2, that the performance with the binary weights is generally somewhat better than that using the normal term frequency weights, although statistically, the differences in performance are not significant.

---

* A recall-precision graph such as that of Fig. 1 is obtained by matching queries and documents (using a cosine coefficient), and ranking all documents in decreasing order of query-document similarity. Precision values are then computed at fixed recall levels of 0.1, 0.2, 0.3, etc., for each query, and the resulting values are averaged for a given query set. When recall-precision graphs for different indexing or search methods are shown in the same figure, the curve closest to the upper right-hand corner (where recall and precision are both near 1) reflects the better performance. [3]

The conclusion is that term frequency weighting may be useful under some circumstances, but that it cannot be guaranteed to perform well in all environments.

A second, somewhat different theory, described by Sparck Jones [1], postulates that high frequency terms may be useful to enhance the recall performance, but that matches between query and document terms which occur rarely in a collection of documents should be treated as more valuable than those which occur frequently. A weighting system is then proposed in inverse document frequency (IDF) order which emphasizes terms with low document frequency. Specifically, if d_i is the document frequency of term i, and N the total number of documents, the inverse document frequency weight J_i is defined as

$$J_i = f(N) - f(d_i) + 1 \qquad (2)$$

where $f(x) = y$ for $2^{y-1} < x \leq 2^y$. [1] The function of equation (2) emphasizes terms with low document frequency d_i, that may be expected to be reasonably specific. A weighting system in inverse document frequency order may then be expected to improve the search precision.

The output of Fig. 3 shows that for the collection of 450 documents in medicine, the inverse document frequency weights produce substantial performance improvements over the ordinary binary (0 to 1) weights previously given in Fig. 2. Once again, however, the improvements appear to be collection-dependent. In fact, while the graph of Fig. 4

obtained for the 425 documents in world affairs previously used in Fig. 1, shows some slight improvements for the inverse document frequency weighting at the low recall end of the curve, the significance computations reproduced with the figure indicate that overall the performance differences are not statistically significant.*1 Indeed for high recall, the standard term frequency weights are superior.*2

It appears then that the standard term or document frequency weights cannot be relied upon to produce reliable improvements in performance valid for most retrieval environments, and that additional criteria must be sought if a generally useful determination of term values is to be made. At least three different nonlinguistic approaches are possible: one may look at the distribution characteristics of the term or document frequencies (instead of the frequencies alone); one may investigate the peculiar characteristics of individual document collections in an attempt to find optimum methods for each environment; finally, one can examine the special needs of various user populations.

An approach based on the study of frequency distributions is described in the next few sections of this study.

---

*1 Probability values smaller than 0.05 are normally taken to indicate statistically significant differences between the two performance curves; the reverse is true when the probability values exceed 0.05. Significance output is shown for the t-test and the Wilcoxon signed rank test. [3]

*2 A comparison between the standard binary weights and the inverse document frequency weights given in the first two columns of Table 5 reveals an equally uncertain picture, since the binary weights are superior for very low recall and also for high recall with the Cranfield collection, where as the IDF weights are better elsewhere.

2. Term Frequency Distribution Characteristics

Consider the sample term frequency distributions shown for a given term i in Fig. 5. In case 1 (Fig. 5(a)) the term frequency is the same for all documents of the collection. A term with such a frequency distribution cannot be used to distinguish among the documents of a collection. Case 2 (Fig. 5(b)) represents a rare term which occurs in only one document with a given frequency k . Since this term is concentrated in one document alone, very few query-document matching coefficients will be affected by its presence, and the term is not likely to be important for retrieval. Case 3, on the other hand, covers a term exhibiting considerable variations in its frequency of occurrence within the documents. There is a chance that such a term may be of considerable value in a retrieval environment.

The following conjectures may then be made concerning the importance and value for retrieval of various types of index terms:

a) terms with very high total frequency of occurrence are most likely not very useful because they match too many documents, and cannot therefore discriminate between relevant and nonrelevant items;

b) terms with medium total frequency and reasonably skewed distributions can help retrieve an adequate number of relevant documents, and also provide a high matching coefficient, and therefore good retrieval performance, for those items in which the term appears many times; it is likely that these may be powerful terms;

c) terms with very skewed distributions, occurring in only a few documents, will produce matches for few documents, but matching items will exhibit a high query-document correlation, and stand a good chance of being judged relevant; such terms are likely to be useful, but not as important as terms of category b);

d) very rare terms should be considered to be of some importance, because a match between a query and a document — even though rare — will isolate a few documents from the bulk of the remaining ones; there is some indication that the elimination of rare terms leads to decreased retrieval effectiveness; [4]

e) terms with medium or low total frequency and flat distributions cannot be used to retrieve documents precisely; however, they do differentiate the small class of items in which they occur from the remainder; in that sense, these terms are of some importance.

If a retrieval criterion is chosen which insists on a reasonable recall performance together with adequate precision, the terms of most importance will be those of medium total term frequency and somewhat skewed frequency distributions. A ranking of the term categories in decreasing order of usefulness would then provide the sequence

(b, c, e, d, a) .

The term discrimination model introduced previously ranks the terms in exactly that order. [5] Specifically, the discrimination value of a term is a measure of the variation in the average pairwise document similarity which occurs when the 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 will decrease the average document pair similarity. Contrariwise, a poor discriminator increases the interdocument similarity. By computing the average pairwise interdocument similarity (or equivalently the average similarity between each document and some common central document, or centroid) before and after the assignment of each term, it is possible to rank the terms in decreasing order of their discrimination value (in decreasing order of the difference between average pairwise document similarity before and after the term assignment). A mathematical description of the model is given in the appendix.

Table 1 shows a list of the ten best discriminators and the ten worst discriminators for three collections in aerodynamics (CRAN 424), biomedicine (MED 450), and world affairs (Time 425), respectively. It may be seen from the Table that the good discriminators are all reasonably specific terms with well-defined meanings in their environments. Indeed, the good discriminators are not distinguishable in terms of specificity from collection to collection, thus putting into question the often-heard assumption that a "soft" subject, such as world affairs, is somehow different in general properties of the vocabulary from the "harder" subjects like aerodynamics or medicine. The poor discriminators of Table 1 appear to include more general terms, such as "method", "solution", "increase", "party", but a look at the Table does not necessarily distinguish good and bad terms.

The frequency distributions characterizing the various discriminator types do, however, reveal substantial differences. Fig. 6(a) shows the distribution of a typical good discriminator, while Fig. 6(b)

does the same for a poor discriminator. It may be seen that the term with the high discrimination value has much lower total frequency than its counterpart, and that its occurrence characteristics are more concentrated in a few documents of the collection — it occurs between 16 and 20 times in one document, between 11 and 15 times in another, 8 times in a third, and so on, and it occurs once in only 27 documents out of some 400.

The poor discriminator, on the other hand, is very evenly distributed and has high total frequency. In particular, it occurs once in 221 documents out of 400, and twice in 75 others. Obviously, this term could not be used to distinguish among these items by causing some to be retrieved while others are rejected.

No matter what the term weighting system — whether based on term frequencies, on inverse document frequencies, or on discrimination values — each procedure assigns a term value to each term, and induces a ranking among the terms in decreasing order of the corresponding term values. These values and ranks can then be used in at least three different ways in a retrieval environment:

a) terms with low values (low term frequencies, or high document frequencies, or low discrimination values) can be eliminated from consideration as potential index terms for the collection, or they can be removed if already assigned; the corresponding process may be called CUT;

b) the calculated term values may be used as weights — for example by multiplying any existing term weights by the new calculated values; this process may be termed MULT;

c) finally, methods a) and b) can be combined by removing low value terms, and using the caluclated values as term multipliers (CUT + MULT).

The aim of this investigation is to determine the usefulness and potential effectiveness of the interactions between term frequency weighting, inverse document frequency weighting, and discrimination value weighting, and to ascertain whether a single common procedure can be found to operate equally effectively in a number of different retrieval environments.

3. Experimental Results for Three Collections

The following principal questions can be investigated using the three ranking systems (term frequency (TF) order, inverse document frequency (IDF) order, and discrimination value (DISC) order), and the three weighting processes (cut-off of low value terms (CUT), multiplication by the term values (MULT), and the combination of the two (CUT + MULT)):

a) what is the value of removing terms with high document frequency (IDF CUT) as a device to improve the precision performance?

b) conversely, what is the value of using term frequency weights (TF MULT) as a device to improve recall?

c) what relative weighting should one give to rare and to frequent terms; in particular, when an IDF process is used which emphasizes the rare terms, should it be applied to binary index sets with 0-1 weights, or to term vectors weighted in term frequency (TF) order, the latter emphasizing some of the frequent terms?

d) what is a good way for combining frequency-based (TF or IDF) weights with discrimination (DISC) weights; in particular, is it reasonable to utilize a combination of DISC CUT and IDF MULT using the assumption that DISC CUT first deletes poor discriminators, while IDF MULT then emphasizes the remaining rare terms which would otherwise be swamped if the high-frequency poor discriminators were still present?

e) how reasonable is the assumption implicit in some of the well-known indexing test systems (such as Aslib-Cranfield [6]) that a method which operates well for one collection and/or user environment will also operate well in other environments; can one characterize collections or users, as well as the corresponding optimal automatic indexing methods?

To obtain answers to some of these questions, three document collections are used which exhibit very similar relevance characteristics, but different indexing environments. The basic collection characteristics are summarized in Table 2, while the indexing statistics are shown in Table 3. It is seen from Table 2 that the probability that a given document is relevant to a query (generality) is exactly the same for all three collections. Furthermore the collection and query sizes are approximately equal. However, the subject areas are different and include aerodynamics (CRAN), biomedicine (MED), and world affairs (Time), respectively.

The indexing statistics of Table 3 are derived from an automatic "word stem" process, where high-frequency function words ("of", "and", "but", etc.) are automatically deleted, and the remaining texts are automatically reduced to word stems assigned to the documents as index terms. [3] It is seen from the Table that an increasingly greater number of distinct terms is used for CRAN, MED, and Time respectively. However, the average total frequency of the terms is fairly high and approximately equal for CRAN and Time, while it is quite low (6.2) for MED.

To distinguish between the vocabulary characteristics for CRAN and Time, it is necessary to look at the term frequency distributions rather than only at the average occurrence frequency. This is done in Table 4(a) for total frequencies of occurrence, and in Table 4(b) for document frequencies. The middle column for the MED 450 collection confirms that it contains very few high frequency terms; in fact 44 percent of the terms occur only once, while only 1 percent of the terms occur over 60 times in that collection. When document frequencies are used, it is seen that 55 percent of the MED terms are assigned to a single document alone.

At the low frequency end, the statistics for CRAN and Time are not strictly comparable, because terms with a total frequency of 1 were removed from the Time indexing vocabulary prior to the document assignment to avoid too large a vocabulary. However, the data of Table 4 show that the proportion of high-frequency terms is comparable

for CRAN and Time, while more medium-frequency terms are used in Time, and more low frequency terms in CRAN.

The main questions concerning the relative value of term frequency and inverse document frequency weighting can be investigated using the output of Table 5 which gives average precision values at each of ten recall points for the three collections, using IDF MULT, IDF CUT and IDF CUT + MULT, applied to binary vectors and to vectors using term frequency weights. For the IDF CUT runs, the thresholds used for removing terms with high document frequency were 129, 19, and 104 respectively for CRAN, MED and Time (that is, terms with document frequencies greater or equal to the threshold were deleted). This removes 0.50%, 3.7%, and 0.33% of the terms with highest document frequency, accounting for 11.8%, 9.71%, and 11.1% of the total term occurrences, respectively.

In Table 5, a single bar appears next to the precision values to indicate in each case whether the precision is higher for the binary vectors or for the corresponding term frequency vectors. A double bar is used to flag the single result for each collection giving the best performance of the six possible ones shown in the Table. The data of Table 5 reveal the following information:

a) The use of term frequency weights as opposed to binary weights is nearly always justified; in particular, TF weighting is better almost everywhere for Time with its reasonably large medium frequency vocabulary; it is also generally better for MED, and for CRAN in the low and medium recall range; only for the CRAN collection at very high recall are the binary weights superior.

b) The use of inverse document frequency weights is also justified almost everywhere; only for CRAN at very high recall are the standard binary weights (first column of precision figures in Table 5) better than the IDF runs.

c) whereas the results of Table 5 show that both term frequency as well as inverse document frequency weights are important, the results concerning the best procedure to be followed differ from collection to collection. In particular, the pure term deletion system (IDF CUT) is clearly best for MED; for CRAN and Time on the other hand, the combined deletion and weighting system (IDF CUT + MULT) is preferred except at very high recall.

Obviously, the information in Table 5 provides no clue concerning the use of discrimination values and the relative performance of discrimination and inverse document frequency weighting. The relevant data are shown in Table 6 for two IDF procedures, three DISC methods, and one combined run.

In Table 6 a single bar is used to denote a good performance, while a double bar designates very good output. The IDF CUT runs shown in Table 6 use the same thresholds as those previously used for Table 5, removing approximately ten percent of the total term occurrences. For the DISC CUT runs, the threshold is so chosen that all terms with a negative discrimination value (those for which the average similarity between documents is greater after assignment of a term than before) are removed.

The following principal conclusions may be derived from the data of Table 6:

a) both the IDF and DISC weights are effective in addition to the normal term frequency weights to improve retrieval effectiveness;

b) for low-recall high-precision performance the IDF weights are the best bet; for medium recall requirements, the DISC rankings are generally preferred;

c) for the MED collection with its large population of low frequency terms, the DISC precedures are preferred over the frequency (TF or IDF) rankings;

d) the best overall method is different for each collection, but a combined deletion of poor discriminators and weighting in inverse document frequency order (DISC CUT + IDF MULT) is very effective for all collections.

The summarization of Table 7 may help in interpreting the results. For the CRAN collection, the number of distinct index terms appears approximately correct, because the weighting procedures (IDF MULT and DISC MULT) are generally more effective than the term deletion methods (IDF CUT and DISC CUT). Since there are a fair number of low frequency terms (Table 4(a)), IDF MULT would be expected to improve the precision; since there are also a reasonable number of high frequency terms, DISC weights may help in middle and high recall areas:

Use IDF MULT for very low R - high P;

Use DISC CUT and IDF MULT for medium R;

Use IDF MULT or DISC MULT for high R.

The MED collection is distinct from the others in that it exhibits a very large proportion (80 percent) of very low frequency terms; furthermore, the number of distinct terms is very large compared with the total number of terms. This explains why the term deletion methods (CUT) are generally better than the term weightings (MULT). Furthermore, since there are few high frequency terms (whose emphasis is lessened by the IDF methods), one could expect that the discrimination weights, which are more selective in emphasizing or deemphasizing terms, will be preferred:

Use DISC CUT, or DISC CUT + IDF MULT for very low R;

Use DISC CUT and MULT everywhere else.

The Time collection has term characteristics similar to CRAN, explaining why the term weighting procedures (MULT) are generally better than the deletion methods (CUT). However, Time has more medium frequency terms than CRAN, and they normally provide most of the good discriminators. This explains the success of the term frequency (as opposed to the binary) weights for that collection:

Use IDF MULT for low R - high P;

Use DISC CUT + IDF MULT for medium to high R;

Use IDF MULT or DISC CUT + IDF MULT for high R.

A completely unified recommendation applicable to all three collections is obviously not possible. The combination of term frequency and inverse document frequency weights coupled with deletion

of poor discriminators (DISC CUT + IDF MULT) provides a high standard of performance for all collections. For CRAN and Time alone, the inverse document frequencies (IDF MULT) are effective for downweighting the higher frequency terms, whereas for MED which lacks most high frequency terms and exhibits too many low frequency terms, the discrimination deletion methods (DISC CUT, and DISC CUT + MULT) are most effective.

4. Summary

The amount of improvement obtainable by the IDF and DISC processes over the standard term frequency runs can be ascertained by looking at the statistical significance output of Tables 8 and 9. For each pair of processes listed, t-test and Wilcoxon signed rank test probabilities are given; probability values smaller than or equal to 0.05 indicate significant differences in performance. In each case, the better method is identified as either method A better than B (A > B), or B better than A (B > A), and statistically significant differences are identified by boxes in Tables 8 and 9. The average percentage improvement of the better method over the poorer one is also entered in the Tables.

The upper part of Table 8 reveals that the inverse document frequencies used with term frequency weighting are significantly better than the same methods applied to binary vectors, the percentage of improvement ranging from 5 percent for CRAN to 19 percent for Time. The combined IDF and TF weights are also better than the

standard TF weights alone, the improvement ranging from 15 to 19 percent for IDF CUT + MULT.

Table 9 shows that for the MED and Time collections all IDF and DISC weighting methods are significantly better than the standard TF weights. For MED, the greatest improvement is produced by the DISC CUT and DISC CUT + MULT methods (+22% and +23%, respectively). For the Time collection, the improvement produced by the best methods (IDF MULT and DISC CUT + IDF MULT) is more modest (+11%). Only for CRAN, do some of the IDF and DISC weighting schemes produce advantages that are statistically nonsignificant. Even there, the IDF MULT and DISC MULT processes produce in statistically significant improvements of 14 percent and 11 percent respectively.

The foregoing studies show that considerable opportunities exist for achieving improved retrieval effectiveness using comparatively simple, novel automatic term weighting processes. To identify the best possible process in any given environment, a detailed study must be made of term-, collection-, and user characteristics.

References

[1] 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

[2] H.P. Luhn, A Statistical Approach to Mechanized Encoding and Searching of Literary Information, IBM Journal of Research and Development, Vol. 1, No. 4, October 1957, p. 309-317.

[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] C.S. Yang, On Dynamic Document Space Modification Using Term Discrimination Values, to be published in Scientific Report No. ISR-22, Department of Computer Science, Cornell University.

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

[6] C.W. Cleverdon, J. Mills, and M. Keen, Factors Determining the Performance of Indexing Systems, Vol. 1 - Design, Aslib Cranfield Research Project, Cranfield, England, 1966.

Appendix

The Discrimination Value Model

Consider a set of N documents and let each document j be represented by a set of terms (a term vector), $\underline{V}_j$, where $\underline{V}_{ij}$ represents the weight of term i in document j. Let the centroid C of all document points in the collection be defined as the "mean document", that is,

$$\underline{C}_i = \frac{1}{N} \sum_{j=1}^{N} \underline{V}_{ij} .$$

The centroid is then effectively the center of gravity of the document space. If the similarity between pairs of documents k and j is given by the correlation $r(V_k, V_j)$, where r ranges from 1 for perfectly similar to 0 for completely disjoint pairs, the compactness Q of the document space may be defined as

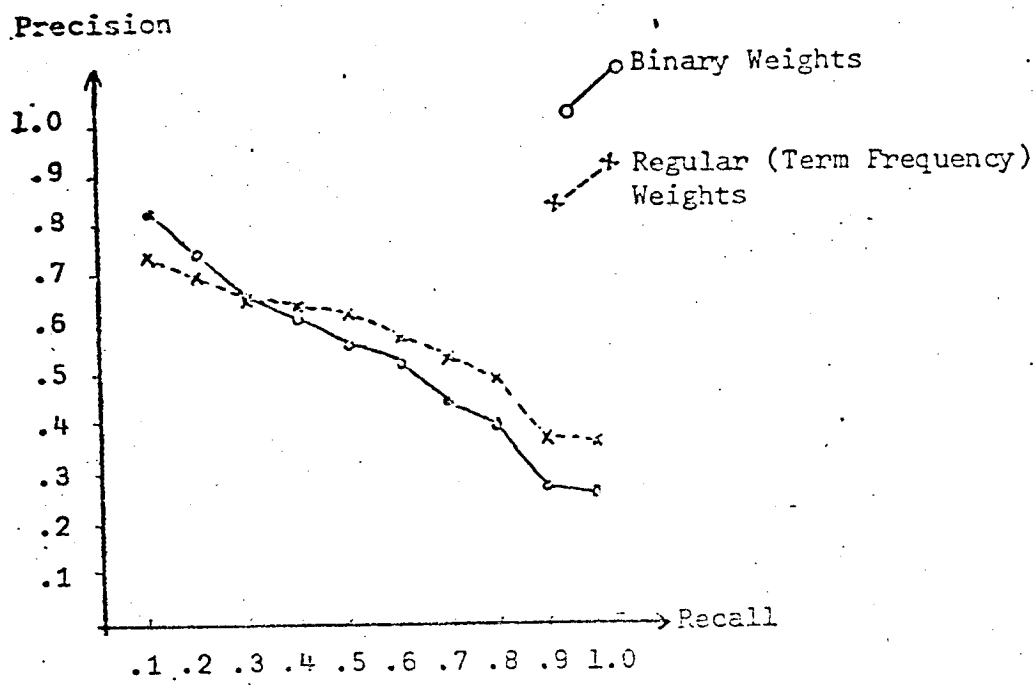
$$Q = \sum_{j=1}^{N} r(C, V_j) , \quad 0 \leq Q \leq N$$

that is, as the sum of the similarities between each document and the centroid. Greater values of Q indicate greater compactness of the document space.

The contribution of a given term m to the space density may be ascertained by computing a function $Q_m - Q$, where Q_m is the compactness of the document space with term m deleted from all document

vectors. If term m is a good discriminator, valuable for content identification, then $Q_m > Q$, that is, the document space after removal of term m will be more compact (because upon addition of that term, the documents will resemble each other less and the space spreads out). Thus for good discriminators $Q_m - Q > 0$. The reverse obtains for poor discriminators for which $Q_m - Q < 0$.

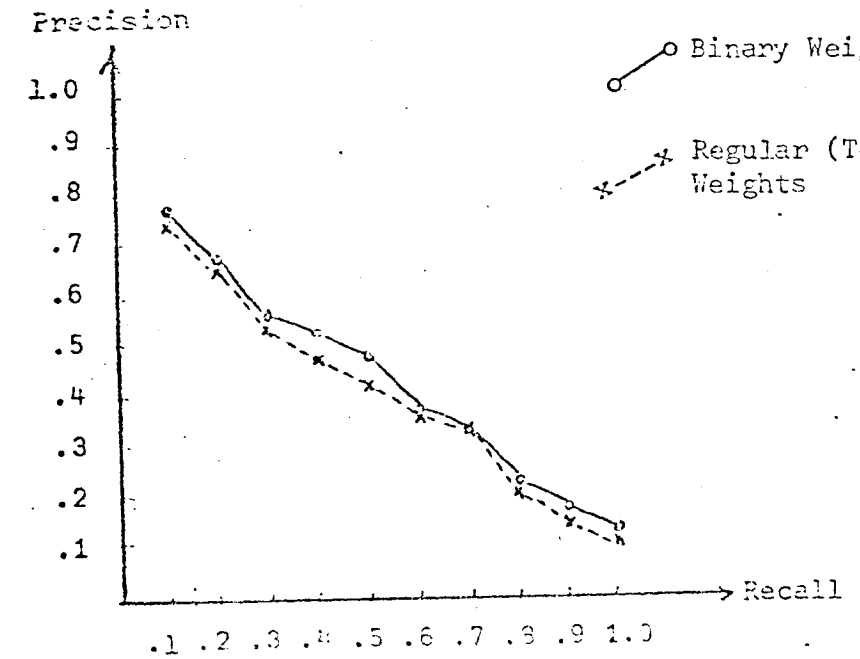
The discrimination value DV_m of term m may then be defined as $Q_m - Q$, and the terms may be ordered in decreasing order of their discrimination value. The number of vector comparisons needed to compute the discrimination values for all M terms is $(M+1)N$, since each individual Q_m value requires N vector comparisons.



R	Precision	
	x---x	o—o
0.1	.7496	.8257
0.3	.6710	.6754
0.5	.6351	.5708
0.7	.5413	.4618
0.9	.3865	.2959
Significance Tests Regular > Binary		
t test :	.0000	
Wilcoxon :	.0000	

Binary vs. Regular (Term Frequency) Weighting
(Time 425 documents, 24 queries)

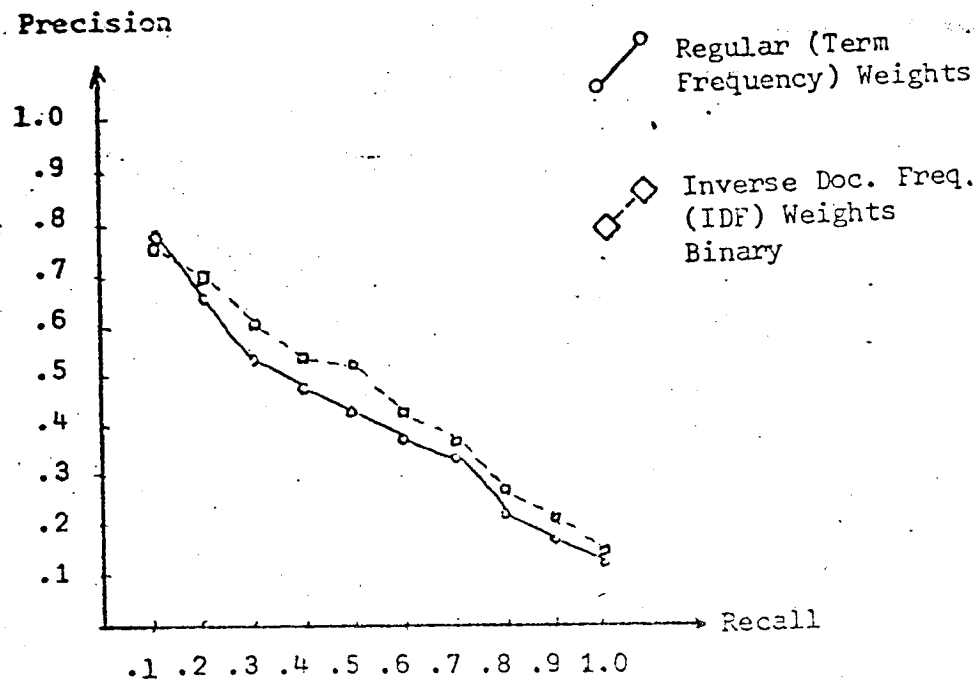
Fig. 1



R	Precision	
	x---x	o—o
0.1	.7891	.7958
0.3	.5481	.5772
0.5	.4384	.4680
0.7	.3357	.3350
0.9	.1768	.1916
Significance Tests Binary > Regular		
t test :	.0626	
Wilcoxon :	.4032	

Binary vs. Regular (Term Frequency) Weighting
(Medlars 450 documents, 24 queries)

Fig. 2

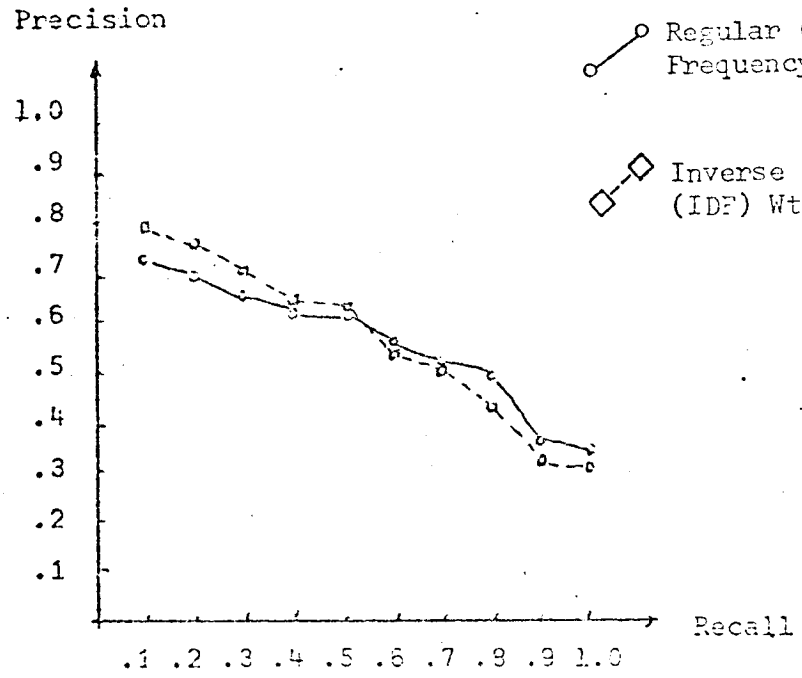


R	Precision	
	o—o	□--□
.1	.7891	.7770
.3	.5481	.6037
.5	.4384	.5315
.7	.3357	.3897
.9	.1768	.2080
Significance Tests IDF > TF		
t test	.0000	
Wilcoxon :	.0000	

Term Frequency vs Inverse Document Frequency Weighting

(Medlars, 450 documents, 24 queries)

Fig. 3

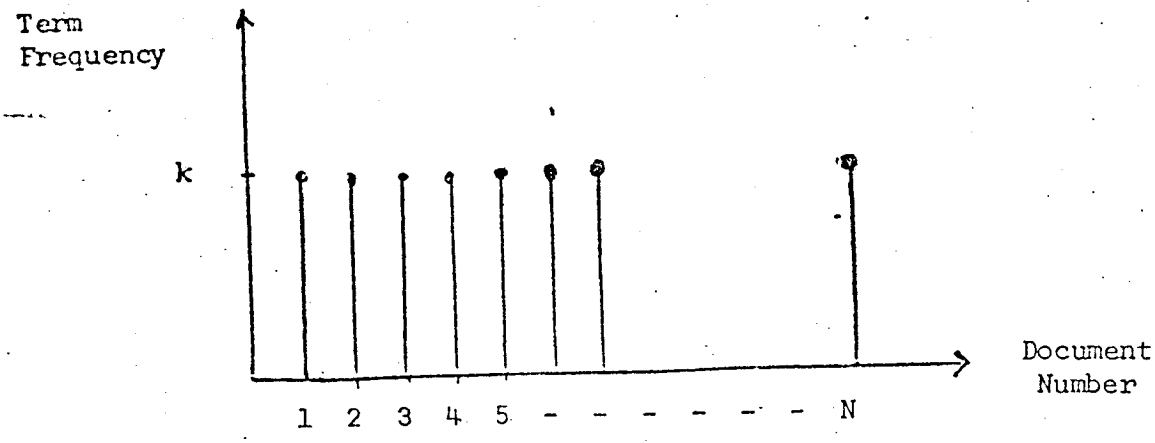


R	Precision	
	o—o	□--□
.1	.7496	.8085
.3	.6710	.7114
.5	.6351	.6219
.7	.5413	.5124
.9	.3865	.3374
Significance Tests IDF > TF		
t test	.2567	
Wilcoxon :	.1240	

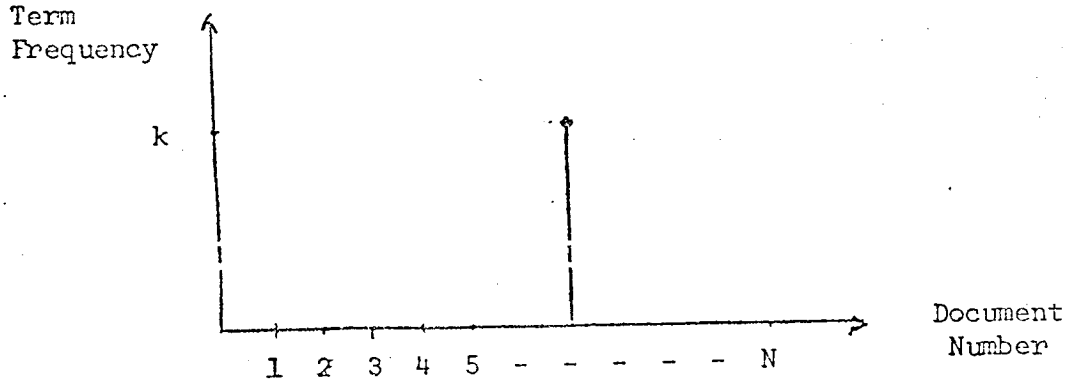
Term Frequency vs Inverse Document Frequency Weights

(Time, 425 documents, 24 queries)

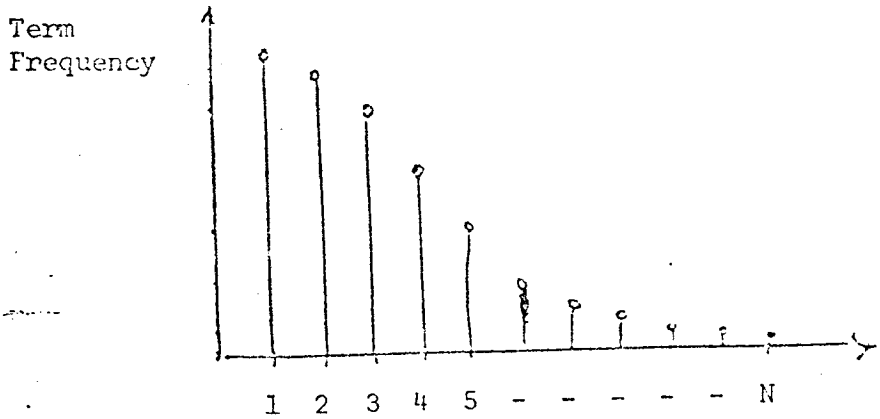
Fig. 4



a) Flat Distribution — Term Occurs with Equal Frequency in All Documents



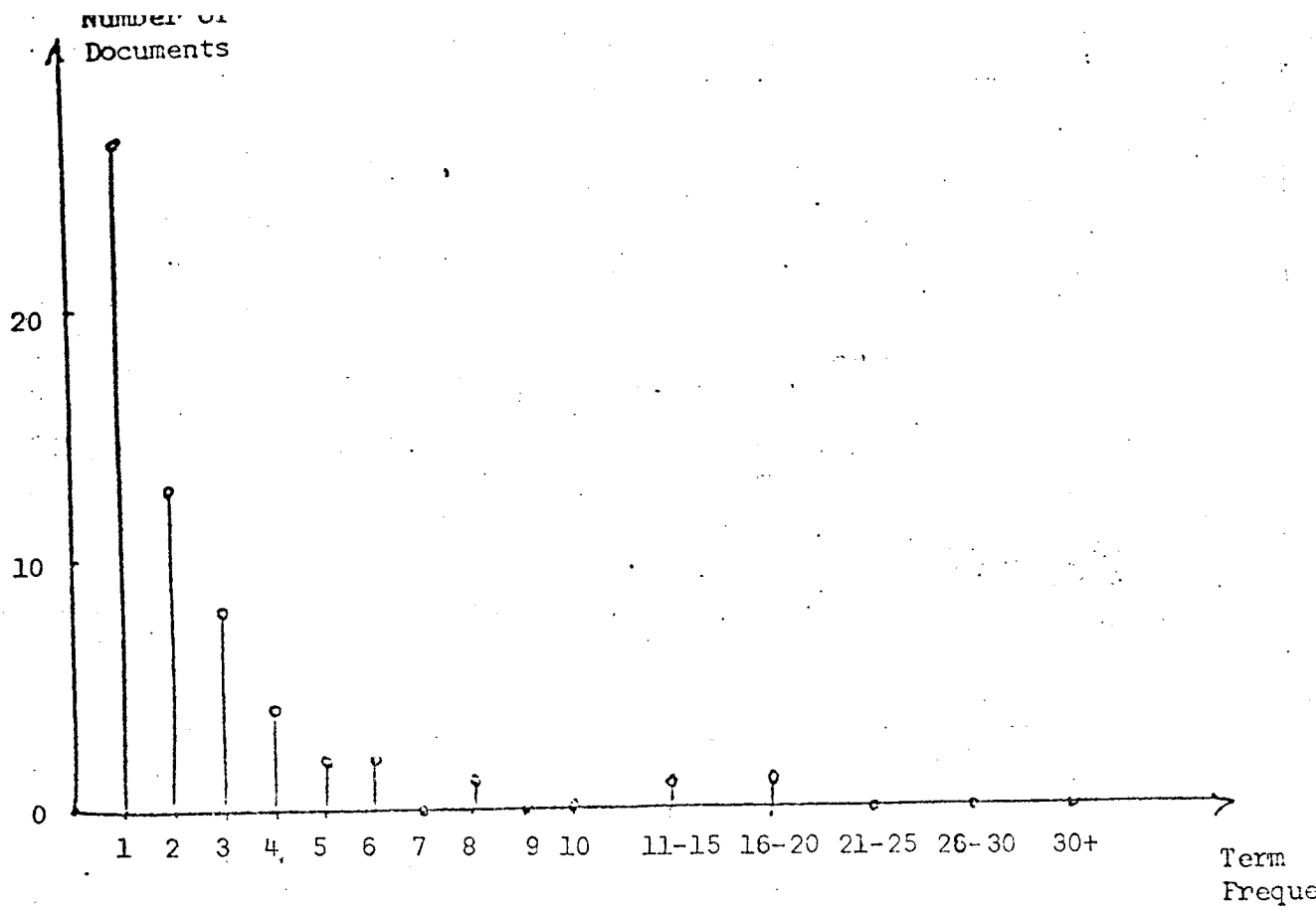
b) Peaked Concentrated Distribution — Term Occurs in Only One Document



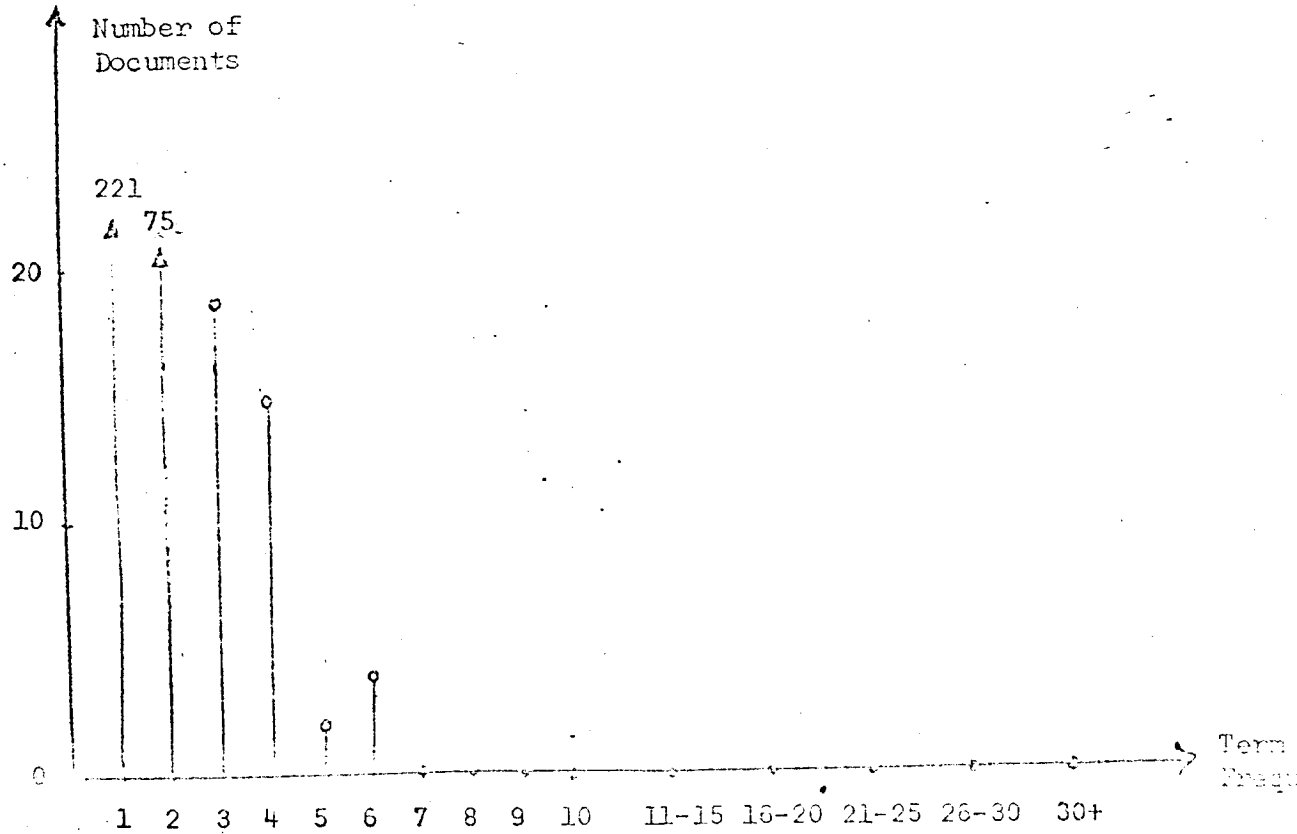
c) Skewed Distribution — High in Some Documents and Low in Others

Some Frequency Distributions for Term i

Fig. 5



a) Good Discriminator (document frequency 59, total frequency 151)



b) Poor Discriminator (document frequency 336, total frequency 522)

Fig. 6

Best Discriminators

Cran 424	MED 450	Time 425
1. Panel	1. Marrow	1. Buddhist
2. Flutter	2. Amyloidosis	2. Diem
3. Jet	3. Lymphostasis	3. Lao
4. Cone	4. Hepatitis	4. Arab
5. Separate	5. Hela	5. Viet
6. Shell	6. Antigen	6. Kurd
7. Yaw	7. Chromosome	7. Wilson
8. Nozzle	8. Irradiate	8. Baath
9. Transit	9. Tumor	9. Park
10. Degree	10. Virus	10. Nenni

Worst Discriminators

Cran 424	Med 450	Time 425
2642 Equate	4717 Clinic	7560 Work
2643 Theo	4718 Children	7561 Lead
2644 Bound	4719 Act	7562 Red
2645 Effect	4720 High	7563 Minister
2646 Solution	4721 Develop	7564 Nation's
2647 Method	4722 Treat	7565 Party
2648 Press	4723 Increase	7566 Commune
2649 Result	4724 Result	7567 U.S.
2650 Number	4725 Cell	7568 Govern
2651 Flow	4726 Patient	7569 New

Best and Worst Discriminators for Three Collections

Table 1

Collection Statistics	CRAN 424	MED 450	Time 425
Subject area	aerodynamics	biomedicine	world affairs
Number of documents	424	450	425
Average document length in words	200	210	570
Number of queries	24	24	24
Relevance count (average number of relevant documents per query)	8.7	9.2	8.7
Generality (relevance count divided by collection size)	0.02	0.02	0.02

Basic Collection Statistics

Table 2

Indexing Statistics	CRAN 424	MED 450	Time 425
Number of distinct terms (word stems)	2,651	4,726	7,569
Total number of term occurrences	35,353	29,193	112,136
Average total frequency of occurrence of terms	14.8	6.2	13.3
Average number of terms per document	83.4	64.8	263.8
Compression percentage of documents (indexing length to word length)	40%	30%	46%

Basic Indexing Statistics

Table 3

Total Frequency of Occurrence	Term Percentage		
	CRAN. 424	MED. 450	Time 425
60 and over	5%	1%	5%
10 to 59	19%	13%	28%
5 to 9	13%	13%	20%
2 to 4	25%	29%	47%
1	38%	44%	-

a) Total Occurrence Frequency Percentages

Total Document Frequency of Terms	Term Percentage		
	CRAN. 424	MED. 450	Time 425
60 and over	3%	0%	2%
10 to 59	16%	8%	24%
5 to 9	13%	12%	20%
2 to 4	25%	27%	45%
1	43%	55%	9%*

*excludes terms of total occurrence frequency equal to 1

b) Document Frequency Percentages

Statistics of Term Occurrences for Three Collections

Table 4

	R	Standard Binary Weights	IDF MULT Binary Weights	IDF MULT TF Weights	IDF CUT Binary Weights	IDF CUT TF Weights	IDF CUT + MULT Binary Weights	IDF CUT + MULT TF Weights
CRAN	.1	.7165	.7502	.7573	.6811	.6975	.7416	.7704
	.2	.5419	.6692	.6241	.5545	.5945	.6644	.6793
	.3	.4581	.5336	.5348	.4832	.5097	.5332	.5574
	.4	.3673	.4146	.4457	.3719	.4197	.4078	.4768
	.5	.3231	.3475	.3935	.3046	.3355	.3390	.3954
	.6	.2664	.2946	.3182	.2536	.2938	.3011	.3213
	.7	.2283	.2431	.2521	.2021	.2326	.2448	.2712
	.8	.2082	.1923	.1953	.1823	.1802	.1925	.2033
	.9	.1538	.1409	.1388	.1335	.1316	.1439	.1402
	1.0	.1439	.1328	.1277	.1215	.1256	.1361	.1306
MED	.1	.7958	.7770	.8459	.7872	.7999	.7874	.8275
	.2	.6912	.7069	.7557	.6692	.7622	.6721	.7548
	.3	.5772	.6037	.6584	.6197	.6865	.5942	.6764
	.4	.5339	.5453	.5442	.5948	.6083	.5708	.5968
	.5	.4880	.5315	.4873	.5299	.5603	.5168	.5457
	.6	.3777	.4179	.4254	.4628	.4682	.4559	.4789
	.7	.3350	.3897	.3833	.4377	.4423	.4200	.4336
	.8	.2421	.2795	.2620	.3084	.3139	.2893	.3066
	.9	.1916	.2080	.2123	.2252	.2452	.2191	.2390
	1.0	.1391	.1490	.1469	.1385	.1524	.1304	.1469
Time	.1	.8257	.8085	.8536	.8306	.8601	.6624	.8975
	.2	.7555	.7741	.7901	.7690	.8268	.6624	.8315
	.3	.6754	.7114	.7568	.7084	.7503	.6441	.7800
	.4	.6224	.6328	.7305	.6164	.7144	.6053	.7574
	.5	.5708	.6218	.6783	.5955	.6872	.6028	.7372
	.6	.5299	.5673	.6243	.5529	.6168	.5438	.6529
	.7	.4618	.5124	.5823	.4737	.5645	.4957	.5912
	.8	.4087	.4384	.5643	.4158	.5017	.4635	.5481
	.9	.2959	.3374	.4426	.3025	.4071	.4445	.4318
	1.0	.2854	.3188	.4170	.2528	.3906	.4445	.4118

Term Frequency (TF) and Inverse Document Frequency (IDF) Experiments

Table 5

		Standard TF Weights	TF Weights IDF CUT	TF Weights DISC CUT	TF Weights IDF MULT	TF Weights DISC MULT	TF Weights DISC CUT + IDF MULT	TF Weights DISC CUT + MULT
CRAN	.1	.6844	.6975	.6654	.7573	.6822	.7179	.6456
	.2	.5303	.5945	.5733	.6241	.6259	.6085	.5708
	.3	.4689	.5097	.5142	.5348	.5446	.5080	.5134
	.4	.3482	.4197	.4654	.4457	.4166	.4859	.4669
	.5	.3134	.3355	.3542	.3935	.3641	.3990	.3719
	.6	.2556	.2938	.2923	.3182	.3075	.3004	.3062
	.7	.1989	.2326	.2341	.2521	.2488	.2361	.2413
	.8	.1631	.1802	.1492	.1953	.1833	.1491	.1534
	.9	.1265	.1316	.1274	.1388	.1348	.1258	.1292
	1.0	.1176	.1256	.1223	.1277	.1279	.1209	.1240
MED	.1	.7891	.7999	.8691	.8459	.7995	.8665	.8322
	.2	.6750	.7622	.8105	.7557	.7255	.7830	.8113
	.3	.5481	.6865	.6677	.6584	.5949	.6579	.6671
	.4	.4807	.6083	.6136	.5442	.5066	.6066	.6230
	.5	.4384	.5603	.5798	.4873	.4530	.5722	.5834
	.6	.3721	.4682	.4912	.4254	.4053	.5047	.5119
	.7	.3357	.4423	.4474	.3833	.3715	.4447	.4690
	.8	.2195	.3139	.2988	.2622	.2460	.2907	.3087
	.9	.1768	.2452	.2325	.2123	.2033	.2262	.2401
	1.0	.1230	.1524	.1459	.1469	.1402	.1468	.1561
Time	.1	.7496	.8601	.7911	.8536	.8406	.8453	.8028
	.2	.7071	.8268	.7485	.7901	.7831	.7657	.7480
	.3	.6710	.7503	.7362	.7568	.7197	.7584	.7286
	.4	.6452	.7144	.7000	.7305	.6901	.7276	.6933
	.5	.6351	.6872	.6777	.6783	.6704	.6891	.6737
	.6	.5866	.6168	.6350	.6243	.6176	.6412	.6347
	.7	.5413	.5645	.5907	.5823	.5727	.5918	.5847
	.8	.5004	.5017	.5510	.5643	.5169	.5601	.5475
	.9	.3865	.4071	.4177	.4426	.4208	.4344	.4259
	1.0	.3721	.3906	.4019	.4170	.4053	.4125	.4085

Comparison of Frequency Weighting and Discrimination Value Weighting

Table 6

Collection	Binary Weights	IDF CUT (TF Weights)	DISC CUT (TF Weights)	IDF MULT (TF Weights)	DISC MULT (TF Weights)	DISC CUT + IDF MULT (TF Weights)	DISC CUT + MULT (TF Weights)
CRAN	Not very effective	Good for low-recall, high-precision	Good for medium recall	Best overall, except for medium recall	Equivalent to IDF MULT except for low recall	Very good for medium recall; not as good as IDF MULT	Not as effective as DISC CUT + IDF MULT
MED	Reasonably effective	Useful for most recall levels	Very good for medium and low recall	Substantial improvement at high recall but DISC is better	Not as effective as IDF MULT	Quite effective for all recall levels especially very low recall	Best overall except at very low recall
TIME	Effective for low recall	Effective for very low recall only	Effective for medium recall	Very effective at high and at low recall	Effective but IDF MULT is better	Best overall especially good for medium recall	Effective but not as good as DISC CUT + IDF MULT
	Use for low low R	Very low R	Medium R	High and low R	Medium R	Low and medium R	Medium to high R

Summarization of Results of Tables 5 and 6

Table 7

	CRAN		MED		Time	
	t-test	Wilcoxon	t-test	Wilcoxon	t-test	Wilcoxon
A. IDF MULT Binary Weights	.1580	.0146	.3126	.4412	.0000	.0000
	B>A		B>A		B>A	
B. IDF MULT TF Weights					+14%	
A. IDF CUT Binary Weight	.0380	.0036	.0000	.0000	.0000	.0001
	B>A		B>A		B>A	
B. IDF CUT TF Weights	5%		6%		14%	
A. IDF CUT + MULT Binary Weights	.0214	.0026	.0000	.0000	.0000	.0000
	B>A		B>A		B>A	
B. IDF CUT + MULT TF Weights	6%		7%		19%	
A. IDF MULT TF Weights	.0000	.0105	.0000	.0000	.0000	.0000
	A>B		A>B		A>B	
B. Standard TF run	+14%		+12%		+11%	
A. IDF CUT TF Weights	.0000	.0105	.0000	.0000	.0000	.0000
	A>B		A>B		A>B	
B. Standard TF run	8%		20%		10%	
A. IDF CUT + MULT TF Weights	.0000	.0000	.0008	.0001	.0000	.0000
	A>B		A>B		A>B	
B. Standard TF run	19%		18%		15%	

Statistical Significance Output for Table 5

Table 8

	CRAN		MED		Time	
	t-test	Wil-coxon	t-test	Wil-coxon	t-test	Wil-coxon
A. IDF CUT (TF Weights)	.0232	.5270	.1215	.4503	.2814	.7910
	A>B		B>A		A>B	
B. DISC CUT (TF Weights)						
A. IDF MULT (TF Weights)	.0075	.0788	.0000	.0000	.0001	.0024
	A>B		A>B		A>B	
B. DISC MULT (TF Weights)			6%		3%	
A. DISC CUT + IDF MULT (TF Weights)	.0329	.0080	.0015	.0101	.0000	.0000
	A>B		B>A		A>B	
B. DISC CUT + MULT (TF wts)	3%		2%		3%	
A. IDF CUT (TF Weights)	.0000	.0105	.0000	.0000	.0000	.0000
	A>B		A>B		A>B	
B. Standard TF run	8%		20%		10%	
A. DISC CUT (TF Weights)	.2841	.6561	.0000	.0000	.0085	.0127
	A>B		A>B		A>B	
B. Standard TF run			22%		8%	
A. IDF MULT (TF Weights)	.0000	.0000	.0000	.0000	.0000	.0000
	A>B		A>B		A>B	
B. Standard TF run	14%		12%		11%	
A. DISC MULT (TF Weights)	.0000	.0000	.0000	.0000	.0000	.0000
	A>B		A>B		A>B	
B. Standard TF run	11%		7%		8%	
A. DISC CUT + IDF MULT (TF Weights)	.0505	.2916	.0000	.0000	.0006	.0001
	A>B		A>B		A>B	
B. Standard TF run			21%		11%	
A. DISC CUT + MULT (TF Weights)	.1296	.4093	.0000	.0000	.0094	.0077
	A>B		A>B		A>B	
B. Standard TF run			23%		8%	

Statistical Significance Output for Table 6