CONTRIBUTION TO THE THEORY OF INDEXING

G. Salton, C.S. Yang, C.T. Yu

TR 73-188

November 1973

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

Contribution to the Theory of Indexing

G. Salton, C.S. Yang, C.T. Yu

Abstract:

An attempt is made to characterize the usefulness of terms occurring in stored documents and user queries as a function of their frequency characteristics across the documents of a collection. It is found that the best terms are those having medium frequency in the collection and skewed frequency distributions. Correspondingly, terms exhibiting either very high or very low document frequency are not as useful.

To improve the indexing vocabulary, it becomes necessary to group low frequency terms into classes, and to break up high frequency terms by forming phrases.

An indexing theory is described based on term frequency considerations, and a new phrase generation method is introduced. The resulting improvements in the indexing vocabulary are evaluated.

1. INTRODUCTION

In information retrieval and automatic text processing, the construction of effective indexing vocabularies has always been considered to be the single most important step. Various theories have been proposed — mostly based on statistical considerations — for the automatic identification of good indexing vocabularies, and for the assignment of appropriate term weights. Among the most important are those due to Luhn, Sparck Jones, and Dennis, respectively. [1,2,3]

The best known of the statistical indexing methods is the one due to Luhn where the value of a term assigned to a document or query is proportional to the term frequency (TF), that is, to the number of times a term occurs in the text of a document or document excerpt. [1] The Luhn theory favors a high recall performance, that is, the retrieval of a large amount of relevant material, since the high frequency terms are essential for this purpose.

A second, complementary theory, described by Sparck Jones, is precision oriented, in the sense that it favors the rejection of nonrelevant material. This is accomplished by treating rare terms as more important for retrieval purposes than frequent ones. Specifically, if the document frequency D_i of term i is defined as the number of documents in a collection in which term i occurs, a weighting function in inverse document frequency (IDF) order can be defined for term i as

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

where N is the number of documents in the collection and $f(x) = \lceil \log_2(x) \rceil$. The function of equation (1) assigns high weights to terms occurring in only a few documents. [2]

It has been shown in earlier experiments that a high standard of performance can be obtained, based strictly on frequency parameters, by assigning to each term occurring in a document or query a weight based on the term frequency multiplied by its inverse document frequency (designated TF · IDF). [4] This function favors terms occurring with high frequency in only a few documents, and presumably leads to high recall output at little loss in precision.

A third frequency-based indexing theory extensively investigated by Dennis, states that raw frequency parameters, such as term or document frequencies, are too coarse to measure the effectiveness of a term, and that the complete frequency distribution of a term across the documents of a collection ought to be used. [3] A typical measure of this type is the signal-noise parameter S_i defined as

$$S_i = \log F_i - \sum_{k=1}^{N} \frac{f_i^k}{F_i} \log \frac{F_i}{f_i^k}, \qquad (2)$$

where f_i^k is the term frequency of term i in document k and F_i is the total term frequency of term i; that is,

$$F_i = \sum_{k=1}^{N} f_i^k .$$

The function of equation (2), like the IDF measure, favors terms exhibiting a high concentration in only a few documents (if term i occurs in only a single document, then $f_i^k = F_i$, and $S_i = \log F_i$), and gives lowest values to those terms with perfectly even distributions (where f_i^k is constant across all N documents).

When frequency-based parameters are used in retrieval, it is found that the effectiveness of the performance varies with the environment. In some instances, the performance of a given parameter is impressive, while in others the results may be mediocre. The term discrimination model has therefore been introduced in an attempt to understand the significance of the frequency parameters for document indexing and content identification.

2. TERM DISCRIMINATION MODEL

In the term discrimination model, a good term is assumed to be one which when assigned as an index term will render the documents as dissimilar to each other as possible; that is, it will cause the greatest possible separation between the documents in the indexing space. [5] Contrariwise, a poor term is one which renders the documents more similar, and therefore makes it harder to distinguish one document from another.

The idea is that the greater the separation of the individual documents, that is, the more dissimilar the respective index term vectors, the easier it will be to retrieve some items while rejecting others; contrariwise, when the documents exhibit similar term vectors, that is when the document indexing space is bunched up, it will be impossible to insure the proper discrimination between relevant and nonrelevant items.

The importance of a term is therefore measured by its discrimination value as follows: Let V_j represent the set of terms (the term vector) assigned to document j, and let V_{ij} measure the weight (e.g. the term frequency) of term i in document j. The centroid of all document points in the collection may then be defined as the center of gravity, or "mean" document C, where

$$C_i = \frac{1}{N} \sum_{j=1}^{N} V_{ij}.$$

If the similarity between pairs of documents k and j is measured by a vector matching function $r(V_k, V_j)$, where r ranges from 1 for perfectly similar to 0 for completely disjoint pairs, the compactness Q of the document space may be expressed as

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

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, and hence greater similarity between the documents.

The contribution of a 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 is then defined as $Q_m - Q$, and the terms may be ordered in decreasing order of their discrimination value.

An examination of the terms of a given collection ranked in decreasing discrimination value order reveals the following characteristics:

a) the best terms, exhibiting the highest discrimination values are those with medium total term frequency F_i, and a document frequency less than one half its term frequency;

b) the next best terms with discrimination values close to zero are those with very low document frequency;

c) the least attractive terms with negative discrimination values are those which have a high document frequency (of the order of the collection size) and a total term frequency exceeding the collection size.

Within each of these categories, terms with relatively flat term distributions (for which the term frequency varies only slightly across the documents of the collection), exhibit lower discrimination values than terms with skewed distributions which occur often in some documents, and rarely in others.

Document frequency characteristics for terms (word stems) falling into the three discrimination value categories are given in Table 1 for three typical document collections of about 450 documents each in aerodynamics (CRAN), medicine (MED), and world affairs (TIME). It may be seen that the class of high frequency, negative discriminators is generally fairly small; however, because of their high individual document frequencies, these terms account for a large proportion of total term occurrences. The class of low frequency terms with discrimination values near zero is normally large, while the number of good discriminators with medium document frequency is smaller in size. For the three sample collections of about 450 documents, the document frequency ranges applicable to the majority of the terms for the three classes of discrimination values are 1-5, 5-30, and 30-160, respectively.

If the discrimination value of a term furnishes an accurate picture of its value for indexing purposes, the situation may then be summarized, as shown schematically in Fig. 1(a). When the terms are arranged in increasing order according to their document frequencies in a collection, the first set of terms with very low document frequency D_i exhibits a discrimination value near zero. The frequency distribution of a typical term in that class is shown in column 2 of Fig. 1(b). Next follow the terms with medium D_i and positive discrimination values; a typical representative is used for column 3 of Fig. 1(b), with a document frequency of 36, and a total frequency of 188 in the collection of 450 documents. The high frequency terms exhibit the poorest discrimination values. The sample term in column 4 of Fig. 1(b) has a document frequency of 337 out of 450 documents; it occurs once in 221 documents, and twice in 75 more items. Obviously, this term is not very useful for distinguishing the documents from each other.

3. CONSTRUCTION OF GOOD INDEXING VOCABULARIES

The discrimination value model described in the previous section suggests the following method for the construction of improved indexing vocabularies: the terms in the low document frequency range must be combined into sets in such a way that the document frequencies of the resulting sets increase; contrariwise, the terms in the high frequency range must be broken up into subsets so as to produce terms with lower document frequency. [6] In terms of the picture of Fig. 1(a), these transformations are designed to move all terms toward the center of the document frequency range, where their effectiveness may be expected to be maximized.

The transformation of type 1 is of course well-known in information retrieval, and consists generally in providing for each term, one or more term substitutes, considered identical for retrieval purposes with the original terms. Specifically, a number of such low frequency terms may be grouped into a given class, or category, and the original term identifiers can then be replaced by the corresponding class identifiers exhibiting higher occurrence frequencies than the original terms. An arrangement of terms into classes represents a thesaurus, and methods have been described in the literature for constructing such thesauruses automatically. [7,8] Furthermore, the extensive evaluation results available for thesaurus utilization indicate that thesauruses normally lead to improved recall performance. This is indicated schematically on the graph of Fig. 1(a).

While the "left-to-right" transformation of low frequency terms into classes of higher frequency terms is important for improved recall, it is even more crucial to effect the reverse "right-to-left" transformation of type 2 of high frequency terms into medium frequency terms, since the high frequency terms are the ones exhibiting the lowest discrimination values.

The classical method for producing lower frequency terms from higher frequency components is to generate "phrases" consisting of several combined terms. For example, in a computer science collection, the terms "program" and "language" may be insufficiently specific, particularly when assigned to a large proportion of the documents in a collection. The phrase "programming language" is more specific and may, when assigned to the documents, lead to improved precision output. Unhappily, whereas a great deal is known about thesaurus construction (term grouping) methods, the experiences obtained with phrase generation procedures have not been uniformly successful. Neither one of the two best known phrase generation methods, involving either the use of syntactic analysis procedures for the formation of phrases, or the use of statistical cooccurrence techniques, has been uniformly satisfactory in retrieval environments. [9]

A new phrase generation system based on the term discrimination model is therefore proposed. Specifically, if the discrimination model outlined in Fig. 1 is in fact an accurate representation of the true situation, it must be possible to improve the retrieval performance by breaking up terms with negative discrimination value in such a way that lower frequency terms are produced from higher frequency components, with correspondingly better discrimination values. Specifically, if the high frequency nondiscriminators are taken in groups, and "phrases" are formed for cooccurring sets of nondiscriminators, the phrases will obviously exhibit lower document frequencies than the original components. The process is illustrated in the example of Fig. 2, for two original high frequency terms T_i and T_j, exhibiting an area of overlap consisting of the documents to which both terms are assigned. The frequency range of T_i and T_j may be reduced, by assigning term T_i' to those documents in which T_i only appears but not T_j; similarly T_j' is assigned to items in which only T_j was originally present, while the phrase T_{ij} is assigned to documents originally containing both terms.

The transformation illustrated in Fig. 2 may of course be generalized by using larger term groups, obtained for example through an automatic term clustering process; these term groups are then broken down into subsets and assigned to the documents and queries in addition to, or instead of, the original high-frequency components. While the proposed phrase formation process neglects syntactic and semantic phrase formation criteria, it is, however, compatible with the discrimination value model which indicates that medium frequency terms with skewed distributions produce effective content indentifiers.

The phrase formation process used for experimental purposes is described and evaluated in the next section.

4. PHRASE GENERATION RESULTS

It was seen in the last section, that phrases are ideally formed from high-frequency components exhibiting substantial cooccurrence characteristics in the documents of a collection. Groups of cooccurring terms are normally obtainable through a term clustering procedure. Term clustering (thesaurus formation) is, however, expensive to undertake; moreover, there is no guarantee that the term combinations (phrases) derived from the term clusters would, in fact, occur with substantial frequency in the user queries, as well as in the documents. However, unless phrases can be assigned to both documents and queries, improvements in performance cannot be obtained.

For this reason, a short-cut process is used experimentally by taking as the initial set of phrase components the set of nondiscriminators occurring in a number of sample user queries. These terms are arranged in increasing order of their discrimination values, that is, worst discriminator first, and groups of threes are formed, as shown in the illustration of Fig. 3.* For each group of three terms, say T_i, T_j, and T_k one triple (T) is formed, denoted as T_{ijk}, as well as three pairs (P), including T_{ij}, T_{ik}, and T_{jk}, respectively. This process is illustrated in Fig. 3.

The retrieval system can now be run in several ways by using the high-frequency nondiscriminators either as single terms (S), pairs (P), or triples (T), or in combination. In each case, a phrase term (pair

*The decision to use initial groups of three terms is a convenient artifice which avoids a term clustering step entirely, while creating term groups and subgroups that are small enough to insure that the phrases will exhibit adequate cooccurrence characteristics in queries and documents.

or triple) is assigned only when all components are present in a given document, or query. The following combinations were tried experimentally:

a) SPT: pairs and triples are added to the original single terms;

b) PT : pairs and triples are added, but the original singles are deleted whenever the corresponding pair or triple is used;

c) ST : triples are added to the original singles;

d) P : pairs are used, and the corresponding singles are deleted.

Results are available for three document collections, consisting of about 450 documents each, in aerodynamics, medicine, and world affairs, together with 24 user queries for each collection. These collections exhibit identical relevance characteristics and have been used earlier for experimental purposes. [4] The collection statistics are shown in Table 2.

For each collection, two sets of single terms are used for phrase formation, consisting respectively of the high-frequency nondiscriminators, and of the medium-frequency discriminators. The reason for using the latter is that their document frequency may still be substantial enough to produce improvements in precision by a phrase generation process. The phrase generation statistics are shown in Table 3. It is seen that 74, 141, and 406 nondiscriminators are used for the three collections, respectively, compared with 587, 661 and 725 discriminators. Following comparison with the query vectors 81, 90 and 96 nondiscriminators are found occurring in user queries, together with 99, 93, and 72 discriminators, for the three collections, respectively.

The illustration of Fig. 3 shows that the number of term pairs generated is always identical to the number of single terms which enter into the process, while the number of term triples is one-third the number of original singles. Table 3 shows that, on the average, 8.4 pairs and 1.35 triples are applied to each document vector for the CRAN collection, compared with the 83.4 single terms per vector used originally. Comparable data for the MED and Time collections are 2.45 pairs and 0.08 triples, and 11.3 pairs and 1.9 triples, respectively.

The frequency distribution of the single terms used for phrase generation is compared with the distributions of the resulting pairs and triple phrases for the three collections in Table 4. It is obvious from the data shown that a much lower average document frequency obtains for the pairs than for the single terms; the document frequency is still lower for the triples than for the pairs. The right-to-left translation from the poor right-hand discrimination region of Fig. 1 to the good central region is therefore certainly taking place.

The corresponding recall-precision results are shown for the three collections in Tables 5 and 6. Table 5 shows average precision values at ten recall points for phrase runs SPT, PT, ST and P; a control run using standard term frequency weighting but no phrases is also included. Results are shown separately for phrases obtained from the high-frequency nondiscriminators and from the medium frequency discriminators. The best results in each section of Table 5 are emphasized by a vertical bar alongside the precision values.

It may be seen from Table 5, that when the high-frequency nondiscriminators are combined into phrases, improvements over the standard TF run are obtained almost everywhere. The best runs are the PT and P runs, where the single term nondiscriminators are deleted when the phrases are introduced into the vectors. Substantial improvements are also obtained for the phrases derived from the discriminators, listed on the right-hand side of Table 5. However, in that case, the good runs are the SPT and ST runs in which the single term discriminators are maintained.*

A combined run in which the phrases obtained from the nondiscriminators are applied using the PT strategy, whereas phrases from discriminators are used with the SPT system is shown in the middle of Table 6, designated as PT + SPT. This phrase procedure is compared against the previously mentioned optimum single term weighting process, labelled TF · IDF (term frequency multiplied by inverse document frequency). The best results are again emphasized by a vertical bar. It is seen that the single term weighting process is somewhat preferable for the CRAN collection; however, the phrase generation methods are superior both for MED and Time.+

The effectiveness of the vocabulary improvement obtained from the phrase generation procedure is summarized by the statistical significance output of Table 7. For each of the three collections the following pairs of runs are compared:

a) term frequency (TF) run against PT phrase run using nondiscriminators;

b) TF run against SPT phrase run using discriminators;

c) TF run against combined PT + SPT; and

d) combined PT + SPT against combined TF · IDF weighting.

*The elimination of the single term nondiscriminators is obviously useful, whereas the elimination of the single term discriminators would bring about considerable losses.

+The TF · IDF weighting system can of course be applied in addition to the phrases.

T-test and Wilcoxon signed rank test probabilities are given for each pair of runs; probability values smaller than 0.05 indicate in each case that the phrase run is significantly better (in a statistical sense) than the standard TF run.

The results of Table 7 show that only for two comparisons using the CRAN collection does the phrase process not perform as expected. In all other cases, the phrase methods produce significant improvements over the standard TF weighting for single terms, and they are also superior to the TF · IDF combined term weighting system.

One may expect that when the "left-to-right" thesaurus transformations are applied to the low frequency nondiscriminators in addition to the right-to-left phrase generation for high frequency terms, an optimal indexing vocabulary results whose effectiveness is unlikely to be surpassable in practical implementations.

REFERENCES

[1] 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.

[2] 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.

[3] S.F. Dennis, The Design and Testing of a Fully Automatic Indexing-Searching System for Documents Consisting of Expository Text, in Information Retrieval — A Critical View, E. Schecter, editor, Thompson Book Co., Washington, 1967.

[4] G. Salton and C.S. Yang, On the Specification of Term Values in Automatic Indexing, Journal of Documentation, Vol. 29, No. 4, December 1973.

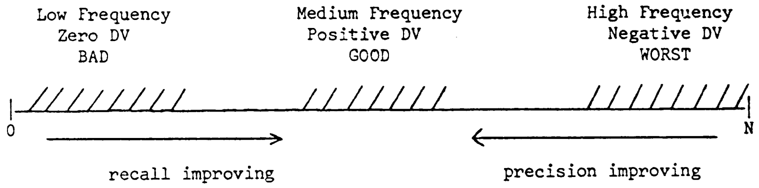
[5] K. Bonwit and J. Aste-Tonsman, Negative Dictionaries, Scientific Report ISR-18, Section VI, Department of Computer Science, Cornell University, October 1970.

[6] C.T. Yu, Theory of Indexing and Classification, Cornell University Doctoral Thesis, Report TR 73-181, Department of Computer Science, Cornell University, August 1973.

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

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

[9] 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.



a) Document Frequency Range

Number of Occurrences f_i of Term i in Collection	Number of Documents with Corresponding Frequency		
	Low Frequency Term Zero DV	Medium Frequency Positive DV	High Frequency Term Negative DV
1	10	26	221
2	3	13	75
3	3	8	19
4	-	4	15
5	-	2	3
6	-	2	4
7	-	-	-
8	-	2	-
9	-	-	-
10	-	-	-
11-15	-	2	-
16-20	-	2	-
21-25	-	-	-
26-30	-	-	-
30+	-	-	-
Total Term Frequency F_i	25	188	527
Total Document Frequency D_i	16	36	337

b) Frequency Distribution for Sample Terms

Characterization of Terms with Varying Discrimination Values

Fig. 1

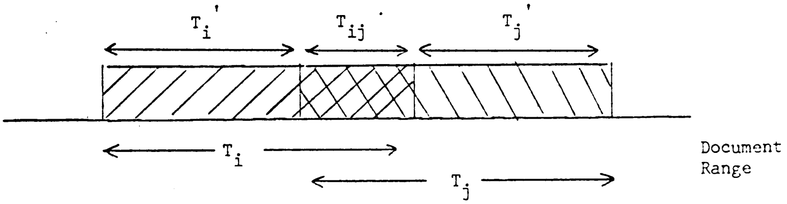
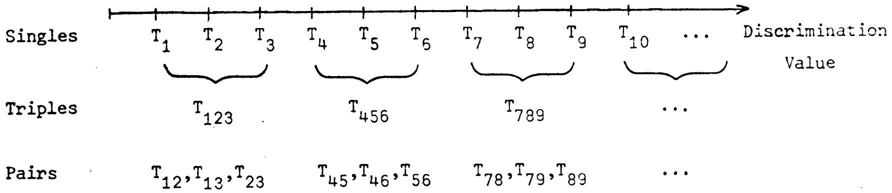


Illustration for Generation of Low Frequency

Term Combinations

Fig. 2


Singles
T_1 T_2 T_3 T_4 T_5 T_6 T_7 T_8 T_9 T_{10} ...
Discrimination Value
Triples
T_{123} T_{456} T_{789} ...
Pairs
T_{12}, T_{13}, T_{23} T_{45}, T_{46}, T_{56} T_{78}, T_{79}, T_{89} ...

Generation of Term Pairs and Triples

from Negative Discriminators

Fig. 3

	Term Characteristics	Low Frequency Terms Zero DV	Medium Frequency Positive DV	High Frequency Terms Negative DV
CRAN 424	Discrimination value range	0 - 0.007	0.007 - 0.254	-2.936 - 0
	Number of terms in range	1990	587	74
	Document frequency range D_i	1 - 10	1 - 67	53 - 214
	Area of concentration of D_i	1 - 5	20 - 40	70 - 160
MED 450	Discrimination value range	0 - 0.008	0.008 - 0.138	-5.025 - 0
	Number of terms in range	3924	141	661
	Document frequency range D_i	1 - 26	1 - 28	14 - 138
	Area of concentration of D_i	1 - 3	5 - 20	20 - 70
TIME 425	Discrimination value range	0 - 0.004	0.004 - 0.247	-1.862 - 0
	Number of terms in range	6468	725	406
	Document frequency range D_i	1 - 39	1 - 63	32 - 271
	Area of concentration D_i	1 - 3	5 - 30	32 - 140

Document Frequency Characteristics for Terms

in Discrimination Value Order

Table 1

Collection Statistics	CRAN 424	MED 450	Time 425
Subject area	Aerodynamics	Biomedicine	World Affairs
Number of documents	424	450	425
Number of queries	24	24	24
Number of distinct term (word stems)	2,651	4,726	7,569
Average document length in words	200	210	570
Average number of terms per document	83.4	64.8	263.8
Average number of term pairs and triples per document	9.75	2.53	13.2
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

Collection Statistics	CRAN 424		MED 450		TIME 425	
	High Freq. Terms (Poor)	Medium Freq. Terms (Good)	High Freq. Terms (Poor)	Medium Freq. Terms (Good)	High Freq. Terms (Poor)	Mediu Freq. T (Good
Number of Distinct Terms in Vocabulary	2651		4726		7569	
Number of Distinct Terms Used for Phrase Generation	74	587	141	661	406	725
Number of Occurrences of such Terms in the Queries*	81	99	90	93	96	72
Number of Newly Created Term Pairs (from T_i, T_j, T_k create T_{ij}, T_{ik}, T_{jk})	81	99	90	93	96	72
Number of Newly Created Term Triples	27	33	30	31	32	24
Average Number of Term Pairs Applied to a Document	7.6	0.8	2.1	0.35	9.3	2
Average Number of Term Triples Applied to a Document	1.2	0.15	0.06	0.02	1.5	0.

*Terms are not distinct, since each term may occur in several queries.

Statistics for Newly Created Term

Pairs and Triples

Table 3

	Document Frequency Range	Single Terms		Term Pairs		Term Triples	
		Poor	Good	Poor	Good	Poor	Good
		Discriminators		Discriminators		Discriminators	
CRAN 424	0	-	-	0	4	1	11
	1-9	0	11	6	81	12	21
	10-19	0	17	20	11	6	1
	20-29	0	18	13	2	2	0
	30-39	0	15	8	1	2	0
	40-49	0	22	6	0	2	0
	50-59	15	7	11	0	1	0
	60-69	5	8	5	0	0	0
	70-79	9	1	2	0	1	0
	80-89	4	0	6	0	0	0
	90-99	4	0	1	0	0	0
	100-129	17	0	3	0	0	0
	130-159	14	0	0	0	0	0
	over 160	13	0	0	0	0	0
MED 450	0	-	-	6	44	14	25
	1-9	0	42	69	48	16	6
	10-19	3	40	13	1	0	0
	20-29	17	11	2	0	0	0
	30-39	33	0	0	0	0	0
	40-49	11	0	0	0	0	0
	50-59	9	0	0	0	0	0
	60-69	8	0	0	0	0	0
	70-79	0	0	0	0	0	0
	80-89	3	0	0	0	0	0
	90-99	4	0	0	0	0	0
	100-129	0	0	0	0	0	0
	130-159	2	0	0	0	0	0
	over 160	0	0	0	0	0	0
TIME 425	0	-	-	0	0	0	3
	1-9	0	2	4	34	9	13
	10-19	0	20	18	28	10	7
	20-29	0	20	17	7	4	1
	30-39	0	9	16	0	6	0
	40-49	8	16	7	2	2	0
	50-59	15	4	7	0	0	0
	60-69	3	1	8	0	1	0
	70-79	8	0	7	0	0	0
	80-89	13	0	3	0	0	0
	90-99	10	0	2	0	0	0
	100-129	7	0	3	0	0	0
	130-159	10	0	0	0	0	0
	over 160	22	0	0	0	0	0

Document Frequency Distribution for High Frequency Nondiscriminators
and for Medium Frequency Good Discriminators
Used in Phrase Generation

Table 4

Collection	Recall	TF Control Run	High Frequency Nondiscriminators				Medium Frequency Discriminators			
			SPT	PT	ST	P	SPT	PT	ST	P
CRAN 424	.1	.6844	.6293	.6620	.6787	.6564	.6917	.4737	.6595	.4582
	.2	.5303	.4797	.5283	.5324	.5404	.5536	.3145	.5087	.2970
	.3	.4689	.4242	.4337	.4694	.4820	.4977	.2740	.4748	.2711
	.4	.3482	.3336	.3430	.3455	.3620	.3787	.2224	.3508	.2106
	.5	.3134	.2903	.3000	.3092	.3106	.3532	.2067	.3134	.1825
	.6	.2556	.2366	.2426	.2529	.2460	.2931	.1697	.2625	.1475
	.7	.1989	.1879	.1942	.1978	.1994	.2176	.1175	.1998	.1152
	.8	.1631	.1572	.1595	.1598	.1590	.1802	.0973	.1617	.0952
	.9	.1265	.1270	.1345	.1272	.1360	.1430	.0813	.1303	.0796
	1.0	.1176	.1198	.1284	.1182	.1299	.1331	.0764	.1217	.0742
MED 450	.1	.7891	.7465	.8609	.8055	.8578	.8223	.6896	.8029	.6896
	.2	.6750	.6705	.7609	.6786	.7652	.7168	.5386	.6733	.5186
	.3	.5481	.5629	.6345	.5587	.6303	.5707	.4529	.5464	.4525
	.4.	.4807	.4999	.5947	.4928	.5905	.5191	.3789	.4767	.3673
	.5	.4384	.4599	.5489	.4497	.5430	.4688	.3242	.4378	.3153
	.6	.3721	.3761	.4889	.3885	.4815	.3807	.2606	.3775	.2606
	.7	.3357	.3371	.4348	.3552	.4370	.3455	.2329	.3411	.2329
	.8	.2195	.2366	.3011	.2273	.3022	.2377	.1469	.2377	.1469
	.9	.1768	.1880	.2033	.1839	.2047	.1985	.1051	.1985	.1051
	1.0	.1230	.1229	.1427	.1213	.1440	.1229	.0914	.1219	.0914
TIME 425	.1	.7496	.7744	.8471	.7545	.8274	.7654	.6307	.7589	.5987
	.2	.7071	.7366	.7952	.7151	.7766	.7654	.6251	.7159	.5712
	.3	.6710	.6708	.7539	.6760	.7586	.7144	.5546	.6853	.5353
	.4	.6452	.6357	.7254	.6431	.7255	.6909	.5017	.6509	.4617
	.5	.6351	.6347	.6732	.6326	.6907	.6644	.4662	.6408	.4377
	.6	.5866	.5859	.6320	.5888	.6363	.6105	.4438	.5922	.4162
	.7	.5413	.5354	.5897	.5482	.5945	.5726	.3987	.5567	.3663
	.8	.5004	.4924	.5320	.5137	.5462	.5355	.3539	.5161	.3263
	.9	.3865	.3996	.3997	.3934	.4038	.4289	.2147	.4069	.2050
	1.0	.3721	.3830	.3862	.3787	.3854	.4155	.1995	.3934	.1911

Average Precision Values at Indicated Recall Points
for Three Collections

Table 5

TF	Standard Term Frequency Weighting (Word Stem Run)
SPT	Single Terms, Pairs and Triples Used in Queries and Documents
PT	Pairs and Triples Used; Corresponding Single Terms Deleted
ST	Single Terms Retained; Triples Added
P	Pairs Added; Corresponding Single Terms Deleted

Collection	Recall	TF Control Run	Best Phrase Process PT + SPT	Best Frequency Weighting TF·IDF
CRAN 424	.1	.6844	.7138	.7573
	.2	.5303	.5929	.6241
	.3	.4689	.5005	.5348
	.4	.3482	.3910	.4457
	.5	.3134	.3485	.3935
	.6	.2556	.2906	.3182
	.7	.1989	.2244	.2521
	.8	.1631	.1798	.1953
	.9	.1265	.1475	.1388
	1.0	.1176	.1394	.1277
MED 450	.1	.7891	.8779	.8459
	.2	.6750	.7846	.7557
	.3	.5481	.6519	.6584
	.4	.4807	.6159	.5442
	.5	.4384	.5692	.4873
	.6	.3721	.4925	.4254
	.7	.3357	.4363	.3833
	.8	.2195	.3015	.2622
	.9	.1768	.2002	.2123
	1.0	.1230	.1427	.1469
TIME 425	.1	.7496	.8860	.8536
	.2	.7071	.7984	.7901
	.3	.6710	.7761	.7568
	.4	.6452	.7461	.7305
	.5	.6351	.7020	.6783
	.6	.5866	.6563	.6243
	.7	.5413	.6010	.5823
	.8	.5004	.5483	.5643
	.9	.3865	.4071	.4426
	1.0	.3721	.3958	.4170

Average Precision Values at Indicated Recall Points

Table 6

TF	Standard Term Frequency Weighting (Word Stem Run)
PT + SPT	Use Pairs and Triples Derived from Nondiscriminators plus Singles, Pairs and Triples Obtained from Discriminators
TF · IDF	Use a Term Weight Consisting of Term Frequency Multiplied by the Inverse Document Frequency

	CRAN 424		MED 450		TIME 425	
	t-test	Wilcoxon	t-test	Wilcoxon	t-test	Wilcoxc
A. Standard TF run vs. B. PT phrases from nondiscriminators	0.18 (A > B)	0.41	0.00	0.00	0.00	0.00
A. Standard TF run vs. B. SPT phrases from discriminators	0.00	0.00	0.00	0.00	0.00	0.00
A. Standard TF run vs. B. Combined PT + SPT phrases	0.02	0.00	0.00	0.00	0.00	0.00
A. TF · IDF weights vs. Combined PT + SPT phrases	0.01 (A > B)	0.00	0.00	0.00	0.78	0.81

Statistical Significance Output for Selected Runs

(Probability that run B is significantly better than run A, except where A > B indicates that test is made in reverse direction)

Table 7