American Journal of Computational Linguistics *Microfiche 55 : 61*

A Comparison of Term Value Measurements for Automatic Indexing

Gerard Salton

Department of Computer Science
Cornell University
Ithaca, New York 14853

This work was supported in part by the National Science Foundation under grant GJ 43505.

Abstract

A number of statistical theories have been proposed capable of identifying individual text words that are most useful for the content representation of written texts and documents. Among these are parameters based on the variance of the word-frequency distribution (NOCC/EK), and on information theoretical (signal-noise S/N) premises. These formal parameters are related to practical automatic indexing techniques--most notably to the discrimination value (DV) method, capable of generating content identifiers (individual words, phrases, and word classes) that distinguish the various texts and documents from each other. It is shown that terms with favorable formal parameters also exhibit desirable semantic characteristics in that such terms are concentrated in documents judged relevant by the respective user populations, and vice-versa for terms with unfavorable formal properties.

1. Theories of Term Importance

Automatic indexing may be considered to be a two-step process: first the automatic identification of linguistic entities useful for the representation of document content, and then the assignment to the prospective content identifiers of weights reflecting their importance for content description. Since these tasks must ultimately depend on a study of the texts or documents under consideration, a great deal can be learned by examining the occurrence patterns of words and other linguistic entities in the documents of a collection. Indeed, among the theories of term importance which have been studied in recent years, the best known ones are based on the respective frequency distributions across a variety of written texts.

A) Variance-Based Measures

The most widely used of the statistical theories distinguishes so-called "specialty" words from "nonspecialty" words by assuming that a deviation from randomness in the occurrence pattern of certain text words is indicative of specialization and hence of good content identifiers. Thus the best content descriptors are terms whose occurrence patterns deviate most strongly from randomness. Since a random sprinkling of the occurrences of a given text word across the documents of a collection leads to word frequency distributions which follow the Poisson model, a comparison of the actual frequency characteristics of a given term with the Poisson distribution leads to the appropriate distinction between good content words and poor ones.

More specifically, since the variance V^k of the frequency distribution of term k is proportional to the total frequency of occurrence F^k for terms whose distribution obeys the Poisson model, a measure of term importance is obtainable by using a formula based on the ratio of V^k to F^k. Some typical formulas used for this purpose are V^k/F^k and $n^2 \cdot V^k/F^k$ where n is the collection size. [1,2,3] The basic mathematical formulations are collected in Table 1.

Formulas	Explanation
n	number of documents in collection
f_i^k	frequency of term k in document i
b_i^k ($b_i^k = 1$ when $f_i^k \geq 1$; $b_i^k = 0$ when $f_i^k = 0$)	binary frequency of term k in document i
$F^k = \sum_{i=1}^{n} f_i^k$	total frequency of term k in collection
$B^k = \sum_{i=1}^{n} b_i^k$	document frequency of term k in collection (number of documents in which the term occurs)
$\bar{F}^k = \frac{F^k}{n}$	average frequency of term k in collection
$V^k = \frac{1}{n} \sum_{i=1}^{n} (f_i^k - \bar{F}^k)^2$ $= \frac{1}{n} \sum_{i=1}^{n} (f_i^k)^2 - (\frac{F^k}{n})^2$	variance of frequency distribution

Basic Frequency Formulas

Table 1

One such variance-based measure used by Dennis under the name of NOCC/EK [3] may be computed as

$$\text{NOCC/EK} = \frac{n}{F^k} \sum_{i=1}^{n} (f_i^k)^2 - F^k. \qquad (1)$$

It is obvious from this formulation that the most effective terms are those whose occurrence frequencies f_i^k in the individual documents deviate strongly from the average frequency F^k/n.

B) Signal-Noise Measure

Another measure based on the characteristics of the frequency distribution of individual text units across the documents of a collection is the signal-noise ratio which varies with the skewness of the frequency distribution. This measure has the form of entropy and assigns the highest value to those terms whose occurrence characteristics exhibit the greatest variation from one document to another; contrariwise low values are assigned to terms with relatively similar frequency patterns in each of the documents of a collection. [3,4] The idea is that terms with even frequency distributions which may occur an identical number of times in each document of the collection cannot be used to distinguish the documents from each other; hence, their assignment for purposes of content representation is counterproductive. The reverse obtains for terms with skewed frequency distributions.

The signal noise value $(S/N)^k$ for term k is defined as

$$(S/N)^k = \log F^k - \sum_{i=1}^{n} \frac{f_i^k}{F^k} \log \frac{F^k}{f_i^k} \qquad (2)$$

The negative term in expression (2) is known,as the noise N^k; it is maximized for even distributions where $f_i^k = F^k/n$ for all f_i^k. The properties of the signal-noise measure are thus very similar to those described earlier for the variance-based formulas.

C) Information Theoretic Considerations

The foregoing development leads to a distinction among the terms in accordance with the relative sizes of the individual term frequencies f_i^k in the documents and the total collection frequency F^k. A question arises about the preferred size of the collection frequency F^k (or of the document frèquency B^k) for terms that are useful as content identifiers. This problem may be tackled by having recourse to certain information-theozetic concepts. Consider the task of supplementing a set of existing index terms identifying a collection of documents by addition of a certain number of new terms. Each new term is then most effective when

a) it provides maximum additional reduction in uncertainty among the documents of the collection (that is, its assignment breaks up existing subsets of documents that cannot be distinguished by the existing term assignments into substantially smaller subsets);

b) it exhibits little redundancy with the previously available terms so that its assignment does indeed optimally divide the various document sets.

The first property is obviously not fulfilled for terms with low document frequency B^k, that is, those assigned to very few documents in the collection, because their assignment provides little additional discrimination among the documents; the second property, on the other hand, does not obtain for terms of high document frequency that may be assigned to a very large number of documents, because such terms will obviously exhibit a good deal of redundancy with the already existing terms.

The conclusion is that the best terms are those whose document frequency B^k, or total frequency F^k, is neither too large nor too small, and whose frequency distribution is skewed in that for some documents f_i^k is much larger than $\frac{F^k}{n}$, and for some others f_i^k is much smaller than $\frac{F^k}{n}$.

D) The Discrimination Value Model

The discrimination value model uses as a point of departure the retrieval capability of the various index terms; specifically, a good content-indicative term is designed to help in the retrieval of material that is wanted (thus enhancing the recall), and in the rejection of material that is extraneous (thus enhancing the precision)*. To produce high recall, that is to retrieve most everything that is relevant, the terms used to identify documents and user queries must be fairly general in nature; high precision, on the other hand, that is the rejection of the nonrelevant material, depends on the use of reasonably specific content identifiers. The indexing problem then reduces to the choice of terms that are specific enough to produce high precision while also being general enough to produce high recall.

In the discrimination value model, the assumption is made that the best terms in this respect are those which cause the maximum possible separation among the documents in the "document space". Consider, in particular, a collection of documents each identified by a set of content identifiers, or index terms. The index term sets for two given documents can be compared to produce a similarity coefficient measuring the closeness between the respective documents.

* Recall is the proportion of relevant material retrieved while precision is the proportion of retrieved material that is relevant. An effective retrieval system is one which produces the highest possible precision for a given level of recall.

The existence of the term sets representing the various documents, and the possibility of computing similarity measures between documents can be used to define a document space for the collection. In such a space two documents appear in close proximity when their similarity coefficient is large; contrariwise, documents exhibiting little similarity are widely separated in the document space. One may then conjecture that a document space which is "bunched up", in the sense that all documents exhibit somewhat similar term sets is not useful for retrieval, since one document cannot then be distinguished from another. On the contrary, a space which is spread out in such a way that the documents are widely separated from each other may provide an ideal retrieval situation since some documents may then be retrieved — hopefully the relevant ones — while others can be rejected.

This suggests that the value of an index term can be ascertained by measuring the amount of spreading in the document space which occurs when that term is assigned to the documents of the collection. Specifically, if Q is the density of the document space without term k present among the content indicators, and Q_k is the density after term k is assigned, then for a good term $Q - Q_k > 0$, since the space will have spread after term k is assigned. Conversely for poor terms $Q - Q_k < 0$.* [5,6] An appropriate

* The density of the space might be computed, for example, as the sum of all pairwise similarities between distinct document pairs, that is

$$Q = \sum_{i \neq j} S(D_i, D_j). \qquad \begin{array}{l} 1 \leq i \leq n \\ 2 \leq j \leq n \end{array}$$

where $S(D_i, D_j)$, $0 \leq S \leq 1$, is the similarity between documents D_i and D_j.

measure of term importance is then the term discrimination value, DV_k, defined as

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

It may be of interest to inquire into the relationship between the discrimination value of a term and the statistical (frequency) parameters introduced earlier. The following conclusions are reached from a study of the indexing vocabularies in several different subject areas, relating the document frequency of a term to its discrimination value: [5]

a) terms with very low document frequency that may be assigned to very few documents in a collection are generally poor discriminators; when the terms are arranged in decreasing order of their discrimination values (where rank 1 is assigned to the best discriminator, rank 2 to the next best, and so on) such terms exhibit ranks in excess of $t/2$ for a total of t existing terms;

b) terms with high document frequencies, comprising those that are assigned to more than 10 percent of the documents of a collection are the worst discriminators, with average discrimination ranks (ranks in decreasing discrimination value order) near t;

c) the best discriminators are those whose document frequency is neither too high nor toc low — with document frequencies between $n/100$ and $n/10$ for n documents; their average discrimination ranks are generally below $t/5$ for t terms.

The vector space analysis then appears to confirm the conclusions derived earlier from the statistical models, that terms which appear in a collection with great rarity or excessive frequency are not optimal for content description purposes.

2. Comparison and Evaluation

The discrimination value analysis can be used to derive an effective indexing policy: since the best terms appear to be those with medium document frequencies, such terms can be directly assigned as content identifiers without further refining transformations. On the other hand, terms with excessively high document frequencies must be made more specific thereby decreasing the frequency of their assignment to the queries and documents of the collection; contrariwise, terms with low document frequencies must be made more general by increasing their assignment frequencies. [5] This can be achieved by joining two or more high frequency terms into term phrases, while assembling a number of low frequency terms into term classes. Obviously, a term phrase exhibits a lower assignment frequency than any phrase component, and vice-versa for a term class which replaces a number of individual class elements.

It was shown earlier that the use of phrases and term classes (thesaurus) constructed in accordance with the frequency requirements imposed by the discrimination value theory produces substantial improvements in retrieval effectiveness (recall and precision). In the present work, additional relationships are examined between the statistical and the vector space models. However, instead of actually using the various term sets in a retrieval environment, an attempt is made to relate the formal frequency and vector spaee properties of the terms to the semantic characteristics of these terms.

Specifically, consider a collection of documents in a given subject area and an appropriate set of user queries pertaining to that area. For each user query, the set of documents can be partitioned into two subsets consisting of the relevant set R and the nonrelevant set I, respectively . Relevance is assumed to be user-specified in such a way that a relevant item is assumed to be one which is related in some sense to the information need expressed by the various user queries. The linguistic, or semantic, character of a given term can now be introduced by assuming that the most valuable content identifiers assigned to a collection of texts are those which are concentrated in the documents specified as relevant to the respective queries, as opposed to the nonrelevant ones, contrariwise, the less valuable terms will be concentrated in the nonrelevant items.

The discussion may be formalized by using the concept of term relevance TR. [7] Consider a term k contained in query Q;. the term relevante TR(k) may be defined as

$$TR(k) = \frac{r_k}{|R|-r_k} \Bigg/ \frac{h_k}{|I|-h_k}, \qquad (4)$$

where r_k and h_k are the number of documents containing term k that are relevant and nonrelevant respectively to query Q, and $|R|$ and $|I|$ are the total number of relevant and nonrelevant documents for that query.* When a term k occurs in more than one query, its term relevance may be taken as the average of the relevance values obtained for the various queries.

* The mathematically undesirable situation when $|R| = r_k$ or when $h_k = 0$ is not likely to occur in a practical environment.

It is clear from the function (4) that high values are assigned to those query terms which are prevalent in the relevant items and rare in the nonrelevant, and vice-versa for those prevalent mainly in the nonrelevant. Furthermore, the terms falling into the former class are likely to be more useful for content representation than those in the latter.

To verify the relationships between the statistical models of word importance and the vector space model, document collections are used in three different subject areas, including aerodynamics (CRAN), medicine (MED) and world affairs (TIME). The vocabularies and user populations are disjoint for these three areas. Results which carry through for all three cases should be extendable to other subject fields as well. The basic collection statistics are contained in Table 2.

It may be seen from the Table that the term relevance is defined for only a relatively small number of terms for each collection, namely 458, 172 and 375 for CRAN, MED, and TIME, respectively. The reason is that a term relevance value is computable only for terms which occur jointly in certain query-document pairs. For small experimental collections operating with a restricted number of queries the size of the corresponding term sets is obviously limited.

Consider now the comparison of the standard statistical term value measures with the term discrimination values obtained by the vector space transformations. Table 3 shows the values of the NOCC/EK and S/N measures (expressions (1) and (2)) obtained for the 50 terms with highest discrimination values and the 50 terms with lowest discrimination values for each of the three test collections. The range of the respective values is given in each case, as well as the average values for each set of 50 terms in percent (that is, on

Characteristics	CRAN 424	MED 450	TIME 425
Subject area	aerodynamics	medicine	world affairs
Number of documents	424	450	425
Number of user queries	155	24	83
Number of terms assigned to collection	2651	4726	7569
Number of terms occurring jointly in queries and document sets	458	172	375

Basic Collection Statistics

Table 2

a scale of 0 to 100). T test values are also shown representing the probability that the two sets of 50 values (for the high DV and low DV terms) could have been derived from a common probability distribution by chance. In statistical significance testing, a t-test value smaller than 0.05 is normally taken to imply a significant difference; that is, the hypothesis that the two sets of values do in fact originate from a common distribution is rejected in such a case. [8]

It may be seen that the ranges of values for the statistical parameters NOCC/EK and S/N exhibit substantial differences for all three collections. The same is true for the corresponding average values. Moreover the differences are in all cases statistically significant. It is then clear that a high discrimination value reflected in the ability of a term to expand the document space upon assignment to the collection also implies favorable statistical parameters in terms of variance and skewed frequency distributions; the converse is true for the low discrimination values.

At the bottom of Table 3, range and average values are given for those terms among the sets of 50 terms for which the term relevance is defined (that is, those which co-occur jointly in some query-document pair). Again the term relevance values are substantially different for the two classes of DV terms, and these differences are statistically significant.

Also included in Table 3 are the multiplicative factors which relate the average values for the 50 high discriminators and the 50 low discriminators for each of the three measures (that is, the factor by which the low average value must be multiplied to obtain the high). It may be seen that this factor is much higher for the term relevance than for either of NOCC/EK or S/N. The actual factors for the term relevance are 6.66, 80.0 and 36.33 for the CRAN, MED, and TIME collections, respectively. This indicates that the high discriminators have very much higher average term relevance than the low discriminators; alternatively expressed, there is substantial agreement between the semantic term relevance concept and the automatically derived term discrimination values.

The data already included in Table 3 are shown in term relevance order in Table 4. The output of Table 4 contains range and average values for NOCC/EK, S/N, and DV for the 50 terms with highest term precision and the 50 terms with lowest precision for the CRAN and TIME collections, respectively. Averages are produced for only 30 high and 30 low precision terms for the MED collection because in the medical environment the small number of available queries (24) made it possible to compute term precision values for only 172 terms in all.

It is clear from the output of Table 4 that the differences in the respective values are substantial in all cases, and the t-test values indicate that they are fully significant. For the three collections under study, ~~the~~ evidence indicates that terms with favorable formal parameters tend to be concentrated in documents identified as relevant by the user population, and vice-versa for terms with unfavorable formal parameters. Also shown in Table 4 are average document frequency ($\bar{B}^k$) and average total frequency ($\bar{F}^k$) values for the high and low relevance terms respectively. It may be seen that the high relevance terms exhibit a much lower frequency spectrum (as expected for good discriminators) than the low relevance terms. Once again, it appears that the term relevance reflecting the semantic properties of the terms in their particular collection environment effects a division among the terms very similar to that obtained by the discrimination value computations.

In earlier work it was shown that the discrimination value theory which leads to the assignment to queries and documents of medium frequency terms (including also phrases constructed from high frequency terms, and term classes made up of low frequency terms) exhibits effective retrieval characteristics. [4,5,6] Typical average retrieval precision values for three different recall levels (recall of 0.1, 0.5, and 0.9) are shown for the three collections in Table 5. The output shows that the use of medium-frequency phrases and term classes improves performance by about 20 percent compared with the assignment of single terms alone. The comparison of Tables 3 and 4 between discrimination values on the one hand, and statistical and semantic parameters on the other, indicates that the same theory which produces such effective retrieval characteristics also conforms to the known statistical and linguistic theories of term behavior.

		50 Terms with High Discrimination Values	50 Terms with Low Discrimination Values
CRAN 424			
NOCC/EK	range	4455 to 925	1599 to 450
	average (in percent)	33.96%	10.96%
	t-test	0.00002	
	average high/average low	3.09	
S/N	range	1.954 to 0.699	1.222 to 0.000
	average (in percent)	60.18%	59.95%
	t-test	0.00002	
	average high/average low	1.00	
Term Relevance TR	range	392.66 to 0.00	74.35 to 0.00
	average (in percent)	14.06% (21 terms only)	2.11% (24 terms only)
	t-test	0.02208	
	average high/average low	6.66	

a) CRAN 424 Collection

Comparison of Statistical Models in

Term Discrimination Values

Table 3

		50 Terms with High Discrimination Values	50 Terms with Low Discrimination Values
<u>MED 450</u>			
NOCC/EK	range	9215 to 1359	7614 to 531
	average (in percent)	29.51%	15.61%
	t-test	0.00002	
	average high/average low	1.89	
S/N	range	2.792 to 0.693	1.738 to 0.126
	average (in percent)	48.46%	23.93%
	t-test	0.00002	
	average high/average low	2.03	
Term Relevance TR	range	874.00 to 0.00	9.43 to 0.00
	average (in percent)	16.0% (12 terms only)	0.20% (24 terms only)
	t-test	0.04274	
	average high/average low	80.0	

b) MED 450 Collection

Comparison of Statistical Models with

Term Discrimination Values (cont.)

Table 3

		50 Terms with High Discrimination Values	50 Terms with Low Discrimination Values
TIME 425			
NOCC/EK	range	13010 to 2330	4712 to 451
	average (in percent)	37.5%	10.81%
	t-test	0.00002	
	average high/average low	3.46	
S/N	range	2.966 to 1.424	1.876 to 0.231
	average	68.85%	26.44%
	t-test	0.00002	
	average high/average low	2.60	
Term Relevance TR	range	2454.00 to 62.62	27.73 to 0.44
	average (in percent)	15.26% (12 terms only)	0.42% (23 terms only)
	t-test	0.03921	
	average high/average low	36.33	

c) TIME 425 Collection

Comparison of Statistical Models with

Term Discrimination Values (cont.)

Table 3

	50 High Relevance Terms $\bar{B}^k$=10.3 $\bar{F}^k$=24.6	50 Low Relevance Terms $\bar{B}^k$=58.9 $\bar{F}^k$=84.0
NOCC/EK	3657 to 420 average 38.95%	1584 to 432 average 20.66%
	t-test 0.00002 average high/average low 1.89	
S/N	1.953 to 0.000 average 42.81%	0.998 to 0.045 average 20.63%
	t-test 0.00002 average high/average low 2.08	
DV	1.223 to 0.002 average 65.52%	0.075 to -1.283 average 25.06%
	t-test 0.00140 average high/average low 2.61	

a) CRAN 424 Collection

Comparison of Term Relevance with

Term Discrimination Values

Table 4

<table>
<tr><th></th><th>30 High Relevance Terms
$\bar{B}^k=9.5$ $\bar{F}^k=24.0$</th><th>30 Low Relevance Terms
$\bar{B}^k=22.5$ $\bar{F}^k=41.9$</th></tr>
<tr><td rowspan="3">NOCC/EK</td><td>2648 to 521
average 48.01%</td><td>2248 to 440
average 36.33%</td></tr>
<tr><td colspan="2">t-test 0.02378</td></tr>
<tr><td colspan="2">average high/average low 1.32</td></tr>
<tr><td rowspan="3">S/N</td><td>1.664 to 0.126
average 61.0%</td><td>1.259 to 0.000
average 46.33%</td></tr>
<tr><td colspan="2">t-test 0.00272</td></tr>
<tr><td colspan="2">average high/average low 1.32.</td></tr>
<tr><td rowspan="3">DV</td><td>0.135 to '0.006
average 62.11%</td><td>0.688 to -1.030
average 56.11%</td></tr>
<tr><td colspan="2">t-test 0.00671</td></tr>
<tr><td colspan="2">average high/averag low 1.11</td></tr>
</table>

b) MED 450 Collection

Comparison of Term Relevance with

Term Discrimination Values (cont.)

Table 4

	50 High Relevance Terms $\overline{B}^k$=12.5 $\overline{F}^k$=45.5.	50 Low Relevance $\overline{B}^k$=94 5 $\overline{F}^k$=164.8
NOCC/EK	13010 to 417 average 16.1%	2266 to 451 average 3.4%
	t-test 0.00002 average high/average low 4.74	
S/N	2.966 to 0.000 average 42.31%	1.371 to 0.126 average 19.25%
	t-test 0.00002 average high/average low 2.20	
DV	0.156 to 0.000 average 94.05%	0.004 to -1.862 average 83.0%
	t-test 0.00148 average high/average low 1.13	

c) TIME 425 Collection

Comparison of Term Relevance with

Term Discrimination Values (cont.)

Table 4

Average Retrieval Precision For Various Recall Levels	CRAN 424	MED 540	TIME 425
A) Low Recall (0.1)			
i) single terms	.6844	.7891	.7496
ii) single terms, phrases and term classes	.8299 (+18%)	.9002 (+12%)	.8398 (+11%)
B) Medium Recall (0.5)			
i) single terms	.3131	.4384	.6351
ii) single terms, phrases and term classes	.4455 (+30%)	.5644 (+28%)	.7006 (+ 9%)
C) High Recall (0.9)			
i) single term	.1265	.1768	.3865
ii) single terms, phrases and terms classes	.1458 (+13%)	.3594 (+51%)	.4821 (+20%)

Recall-Precision Performance for

Medium Frequency Terms

(Discrimination Value Theory)

Table 5

References

[1] A. Bookstein and D.R. Swanson, Probabilistic Models for Automatic Indexing, Journal of the ASIS, Vol. 25, No. 5, September-October 1974, p. 312-318.

[2] D.C. Stone and M. Rubinoff, Statistical Generation of a Technical Vocabulary, American Documentation, Vol. 19, No. 4, October 1968, p. 411-412.

[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 Review, G. Schecter, editor, Thompson Book Co., Washington, 1967, p. 67-94.

[4] G. Salton, A Theory of Indexing, Regional Conference Series in Applied Mathematics No. 18, Society for Industrial and Applied Mathematics, Philadelphia, 1975.

[5] G. Salton, C.S. Yang and C.T. Yu, A Theory of Term Importance in Automatic Indexing, Journal of the ASIS, Vol. 26, No. 1, January-February 1975, p. 33-44.

[6] G. Salton, A. Wong, and C.S. Yang, A Vector Space Model for Automatic Indexing, Communications of the ACM, Vol. 18, No. 11, November 1975, p. 613-620.

[7] C.T. Yu and G. Salton, Precision Weighting — An Effective Automatic Indexing Method, to be published in Journal of the ACM, 1976.

[8] D. Williamson, R. Williamson, and M. Lesk, The Cornell Implementation of the SMART System, in The SMART Retrieval System, G. Salton, editor Prentice-Hall, Englewood Cliffs, NJ, 1971, Chapter 2.