A COMPARISON OF SEARCH TERM WEIGHTING: TERM RELEVANCE vs. INVERSE DOCUMENT FREQUENCY

Harry Wu and Gerard Salton*
Cornell University

Abstract

The term relevance weighting method has been shown to produce optimal information retrieval queries under well-defined conditions. The parameters needed to generate the term relevance factors cannot unfortunately be estimated accurately in practice; futhermore, in realistic test situations, it appears difficult to obtain improved retrieval results using the term relevance weights over much simpler term weighting systems such as, for example, the inverse document frequency weights.

It is shown in this study that the inverse document frequency weights and the term relevance weights are closely related over a wide range of the frequency spectrum. Methods are introduced for estimating the term relevance weights, and experimental results are given comparing the inverse document frequency with the estimated term relevance weights.

1. Introduction

The term relevance weight, also known as term precision, is defined as

$$w_i = \log\{\frac{p_i}{1-p_i} \div \frac{q_i}{1-q_i}\} \qquad (1)$$

where p_i represents the probability of occurrence of term i in a relevant document, and q_i is the probability of occurrence of term i in a nonrelevant document. In practice it is convenient to replace the probabilities by frequencies, in which case the relevance weight becomes

$$w_i = \log\{\frac{r_i}{R-r_i} \div \frac{s_i}{I-s_i}\}, \qquad (2)$$

where r_i and s_i represent respectively the number of relevant and nonrelevant documents containing term i, and R and I represent the total number of relevant and nonrelevant items in the collection with respect to the query under consideration. The weighting function of expression (2) is defined simply as the logarithm of a particular proportion of relevant items containing a given term divided by the same proportion of nonrelevant documents containing the term.

It has been shown that the term relevance factor is an optimal query weighting system under the following conditions [1-4]:

i) the terms are assigned independently to the document of a collection;

ii) a binary indexing system is used for the documents; that is, a term is either assigned or it is not assigned to a given document;

iii) the similarity between a query and a document is computed as the inner product of the corresponding term vectors (that is, assuming document $D = (d_1, d_2, \ldots, d_t)$ and query $Q = (q_1, q_2, \ldots, q_t)$, the similarity is defined as

$$s(D,Q) = \sum_{i=1}^{t} q_i d_i).$$

Unfortunately the optimality result of the term relevance is of no consequence unless methods

*Department of Computer Science, Cornell University, Ithaca, New York 14853. This study was supported in part by the National Science Foundation under Grant IST-79-05301.

exist for estimating the term occurrence probabilities p_i and q_i, or alternatively the term frequencies r_i and s_i, as well as the collection statistics R and I. This information is easily generated only when exhaustive relevance assessments are available for each document with respect to each query. In practice such relevance assessments are of course not available before a search is actually conducted. Even after conducting an initial search effort, relevance information is normally obtainable for only a few of the retrieved documents.

In the experiments conducted so far with the term relevance weights various assumptions have been used, leading to the generation of estimated term relevance values. The following main approaches may be cited:

i) The available document collection is broken down into two halves, known as the even and odd collections, respectively; a single set of user queries is used first with the even collection to generate relevance information and compute the term relevance weights for the query terms. The weighted queries are then processed against the odd collection to test the effectiveness of the relevance weighting system. [5-8]

ii) The previous method represents a type of relevance feedback where relevance information obtained from a portion of a collection is used to improve the retrieval characteristics of the remaining documents. However the process is realistic only in selective information dissemination (SDI) situations where the same queries are repeatedly processed against many different collections. A possibly more realistic testing situation keeps the document collection intact, but breaks a query collection into two pieces (the even and odd query sets). The relevance weights are then computed using the even users queries and later applied to new users represented by the odd query set. [9,10]

iii) An even simpler approach consists in noting that R, the number of relevant documents with respect to a query is necessarily very small, compared to N, the total number of items in a collection. From this it follows that $p_i \approx$ constant; $N \approx I$; and $f_i \approx s_i$, where $f_i = r_i + s_i$ represents the frequency of occurrence of term i in the collection. [11] In these circumstances, the term precision formula of expression (2) reduces to

$$w_i = \text{constant} + \log \frac{N-f_i}{f_i}. \qquad (3)$$

iv) Another possibility consists in taking into account dependencies between the terms assigned to the document collection. In that case, the user queries can be modified by adding dependent terms to the originally available query terms. The relevance weighting formulas (expressions 1 and 2) must then be appropriately modified by including the dependent term information. [12]

v) A last possibility for estimating the value of r_i is to assume that a functional relationship exists between f_i, the total number of items in which a term occurs and r_i, the total number of relevant items for the terms. [13,14] This method forms the basis for the experiments described in the present study. The details are therefore covered in the remainder of this report.

It is clear that the use of term relevance weights raises problems of principle and procedure that do not arise in many other term weighting situations. Furthermore, the term relevance weights do not automatically produce large-scale improvements in retrieval effectiveness when compared with other less sophisticated weighting methods. One easily computable term weighting system that has consistently given excellent retrieval results is the so-called inverse document frequency (IDF) method where the weight of a term is inversely related to the number of documents to which the term is assigned. [15]

The evidence available so far indicates that the term relevance weights generated by the previously described estimation methods do not produce retrieval results that are substantially better than IDF. The approximation process discussed under iii) above indicates moreover that under certain simplifying assumptions the term relevance of expression (2) reduces to a form of the inverse document frequency. (The weighting function of expression (3) increases as the document frequency f_i decreases.)

In the remainder of this study, the relationship between term relevance and inverse document frequency are explored in detail and evaluation results are given to demonstrate the differences between the two term weighting methods.

2. Term Relevance and Inverse Document Frequency

A) Term Occurrence Characteristics

To use the term relevance weights of equation (2) it is necessary to determine the values of R and I, that is, the total number of relevant and nonrelevant items in the collection with respect to a given query (R + I = N). If exact values are not available for these parameters, one might define R as the average number of relevant items in the collection for a set of previously processed queries, or R could be taken simply as the set of relevant documents which the user wishes to retrieve with respect to a given query.

It remains to determine the parameters r_i and s_i for each term of frequency f_i ($f_i = r_i + s_i$), that is, the number of relevant and nonrelevant documents in which a given term of frequency f_i may be expected to occur. (The subscript i is dropped in the following discussion.) In choosing terms for incorporation into a query, a given user is unlikely to use the perfect terms, that is, those occurring only in the relevant documents and in none of the nonrelevant ones. On the other hand, the user may be expected to do better than picking merely average terms that are randomly sprinkled among the relevant items in a collection.

Consider the frequency picture of Fig. 1 relating the total frequency of occurrence f of a given term with the frequency of occurrence, r, of the term in the relevant documents. A <u>random</u> term of total frequency f may be expected to occur in a fraction (R/N)f of relevant documents; that is, r = (R/N)f corresponding to line OTB in Fig. 1. A <u>perfect</u> term, on the other hand occurs only in the relevant documents assuming f ≤ R. That is, for perfect terms r = f when $0 \le f \le R$, and r = R for $R \le f \le N$. This corresponds to lines 0A and AB in Fig. 1.

Following (14) it may be reasonable to assume that the behavior of terms actually chosen by the user for incorpoation in a query falls somewhere between the perfect and the random cases. In particular, the assumption is made that the number of relevant documents in which a term occurs is relatively larger for low-frequency than for high-frequency terms, that is,

$$\text{i) } r = af \text{ for } 0 \le f \le R \text{ where } R/N < a < 1 \text{ and}$$

$$\text{ii) } r = b + cf \text{ for } R \le f \le N \text{ where } 0 < c < R/N. \quad (4)$$

The behavior of the actual query terms thus corresponds to the dashed lines 0M and MB of Fig. 1. The parameters a and c represent the slope of lines 0M and MB respectively.

Using these assumptions, it is now possible to characterize the behavior of the term relevance weight of expression (2).

B) Characteristics of Term Relevance

<u>Theorem 1</u>:

For $f \le R$

$$w(f) \approx \log\left(\frac{a}{1-a}\right) + \log\left(\frac{I}{R}\right) - \frac{(1-a)f}{I} + \sum_{i=1}^{\infty} \frac{1}{i}\left(\frac{af}{R}\right)^i . \quad (5)$$

Proof: For $f \le R$, $r = af$

Hence $s = f - r$

$= (1-a)f$.

Therefore

$$r/s = a/(1-a). \quad (6)$$

By definition $w(f) = \log\left(\frac{r}{R-r} \div \frac{s}{I-s}\right)$

$= \log\left(\frac{r}{s} \cdot \frac{I}{R} \cdot \frac{I-s}{I} \cdot \frac{R}{R-r}\right)$

Substituting (6) one obtains

$$w(f) = \log\left(\frac{a}{1-a}\right) + \log\left(\frac{I}{R}\right) + \log\left(\frac{I-s}{I}\right) - \log\left(1-\frac{r}{R}\right). \quad (7)$$

By the normal series expansion of the logarithm, one has for $(u) < 1$

$$-\log(1-u) = u + \frac{u^2}{2} + \ldots + \frac{u^i}{i} + \ldots + = \sum_{i=1}^{\infty} \frac{u^i}{i} \quad (8)$$

By substitution into (7) one obtains

$$w(f) = \log\left(\frac{a}{1-a}\right) + \log\frac{I}{R} - \sum_{i=1}^{\infty} \frac{1}{i}\left(\frac{s}{I}\right)^i + \sum_{i=1}^{\infty} \frac{1}{i}\left(\frac{r}{R}\right)^i \quad (9)$$

Since $s \le f \le R \ll I$, s/I is very small and thus the higher order terms of s/I can be ignored and the theorem is proved with the replacement of s by (1-a)f and r by af.

When a is small, or when $f \ll R$, af/R is also very small and the higher order terms of af/R can also be dropped. One obtains the following corollary:

Corollary 2: When a is small, or when f << R then

$$w(f) \approx \log(\frac{a}{1-a}) + \log(\frac{I}{R}) - \frac{(1-a)f}{I} + \frac{af}{R} \qquad (10)$$

$$= \text{constant}_1 + \text{constant}_2 \cdot f \ .$$

From the corollary it can be seen that w is approximately linear when $0 < f \leq R$ and when a is not loo large. However as a and f increase, the higher order terms of f are required. As $a \to 1$ and $f \to R$, the sum $\sum_{i=1}^{\infty} \frac{1}{i}(\frac{af}{R})^i \to \infty$ indicating that the relevance weight w itself goes to infinity.

The next theorem studies the behavior of the term relevance function in the region $R < f \leq N$. The rate of decrease of w in that region is shown to be approximately the same as the rate of decrease of log(1/f).

Theorem 3: For $f > R$

$$w(f) = \log(\frac{r}{f}) + \log(\frac{I}{(1-a)R} - 1) + \sum_{i=1}^{\infty} \frac{1}{i}(\frac{r}{f})^i \qquad (11)$$

To carry out the proof the following technical lemma is required:

Lemma 4:

For $f > R$,

$$\frac{I-s}{R-r} = \frac{1}{(1-a)}\frac{I}{R} - 1. \qquad (12)$$

The proof is included in the appendix.

The main theorem may now be proved easily.

Proof: For $f > R$

$$w = \log \frac{r}{s}\frac{I-s}{R-r}$$

$$= \log \frac{r}{f-r} + \log(\frac{1}{1-a}\frac{I}{R} - 1) \qquad \text{from (12)}$$

$$= \log \frac{r}{f}\frac{f}{f-r} + \log(\frac{1}{1-a}\frac{I}{R} - 1)$$

$$= \log \frac{r}{f} - \log(1-\frac{r}{f}) + \log(\frac{1}{1-a}\frac{I}{R} - 1)$$

$$= \log \frac{r}{f} + \sum_{i=1}^{\infty} \frac{1}{i}(\frac{r}{f})^i + \log(\frac{1}{1-a}\frac{I}{R} - 1) \ .$$

QED

The last transformation again uses the series expansion (8) for the logarithm.

When $f > R$, $r/f \leq R/f < 1$. As f increases, r/f becomes very small. The theorem can then be further reduced as follows:

$$w(f) \approx \log(\frac{r}{f}) + \frac{r}{f} + \log(\frac{I}{(1-a)R} - 1). \qquad (13)$$

Expression (13) represents the first order case corresponding to (11). Furthermore, since $r \leq R < f$, $|\log(r/f)| >> r/f$.

$$w(f) \approx \log(\frac{r}{f}) + \log(\frac{I}{(1-a)R} - 1). \qquad (14)$$

By the assumptions leading to the frequency spectrum of Fig. 1, r is increasing at a much slower rate than f when $f > R$. Expression (14) then shows that the rate of decrease of w is determined mainly by the rate of decrease of log(1/f). That is, for large f, the term relevance weight decreases at approximately the same rate as the inverse document frequency. As the next theorem shows, the connection between the two weighting systems is even stronger: Not only is the rate of decrease similar for the two systems, but in fact the corresponding values themselves are similar.

Consider first an intuitive argument illustrated in Fig. 2. Assuming $a \approx 1/2$, the value of r is approximately equal to R/2 at f = R. Since r = R at f = N, the rate of increase of r between R and N will be $R\frac{R/2}{N-R}$. The value of R is normally much smaller than N. Hence the rate of increase of r will be very small, indicating that over a wide range of f, the value of r remains approximately equal to R/2. In other words, for a certain range of the frequency f, $r \approx R-r$. From expression (2) for the term relevance, it follows that $w \approx \log((I-s)/s)$. When $N >> f >> R$, one has $N \approx I >> f \approx s >> R \Rightarrow r$. That is, I-s will be close to N, and f close to s. Hence in that case

$$w(f) \approx \log N/f = IDF(f) \ .$$

This relationship is formalized in the next theorem.

Theorem 5: Given $\epsilon > 0$, if $\frac{1}{2+\epsilon} < \frac{r}{R} < \frac{1+\epsilon}{2+\epsilon}$

then

$$|w(f) - IDF(f)| < |\log(I/N)| + 2\epsilon. \qquad (15)$$

The conditions of the theorem imply that r/R is not too far removed from 1/2 as stated earlier.

Proof: Assume f > R

$$w = \log\left(\frac{r}{s} \cdot \frac{I-s}{R-r}\right) = \log\left(\frac{r}{f-r} \cdot \frac{I}{R} \frac{(1-s/I)}{(1-r/R)}\right)$$

$$= \log\left(\frac{N}{f} \frac{r}{R((f-r)/f)} \frac{I}{N} \frac{1-s(s/I)}{1-(r/R)}\right)$$

$$= \log\left(\frac{N}{f} \frac{1}{(1-r/f)} \frac{r}{R} \frac{(1-s/I)}{(1-r/R)} \frac{I}{N}\right)$$

$$= \log\left(\frac{N}{f}\right) - \log\left(1-\frac{r}{f}\right) + \log\left(1-\frac{r}{R}\right)) - \log\left(1-\left(1-\frac{r}{R}\right)\right)$$

$$+ \log\left(1-\frac{s}{I}\right) + \log\left(\frac{I}{N}\right) .$$

Hence

$$w\text{-IDF} = \log\left(\frac{I}{N}\right) - \log\left(1-\frac{r}{f}\right)$$

$$+ \log\left(1-\frac{s}{I}\right) + \log\left(1-\left(1-\frac{r}{R}\right)\right) - \log\left(1-\frac{r}{R}\right). \quad (16)$$

To prove the thereom, it suffices to show that the sum of the last four terms of (16) is less than 2ϵ. This is done by Lemma 6 proved in the appendix.

The conditions of theorem 5 may be illustrated by taking a typical case and exhibiting the corresponding error bounds. Table 1 contains the corresponding data for a small collection where N = 1020, R = 20, I = 1000 and a = 0.53. It is seen from the Table that as the error ϵ increases from 0.5 to 1 and finally to 2, the frequency range for which the IDF approximation holds increases substantially. The theorem shows that the inverse document frequency weight is closest to the relevance weight at medium frequencies. The error increases for very low and very high frequencies as suggested previously in the illustration of Fig. 2.

For the case under consideration, the error is less than 1.02 when f is between 90 and 221. When the range is expanded to $51 \leq f \leq 310$, the error bound increases to 2.02. Finally, the error bound reaches 4.02 when the range is extended to include frequencies between 31 and 509.

The experimental output included in the next section shows that the error bounds stated in theorem 5 are much larger than the real errors likely to be found in practice. In other words the actual similarity between term relevance and IDF will be larger than can be inferred from the theorem.

3. Experimental Output

It was shown in the previous section that for medium values of a, that is, for a near 1/2, the term relevance weight can effectively be approximated by an inverse document frequency weight for term frequencies that are neither too small nor too large. To obtain a better idea of the usefulness of the IDF approximation, actual values are calculated for the case illustrated in the previous section (N = 1020, R = 20, I = 1000) for three different values of a (a = 0.25, a = 0.53, and a = 0.75). Four different weighting formulas are used experimentally:

i) The actual term relevance (expression (2))

$$w_1(f) = \log\left(\frac{r}{f-r} \cdot \frac{I-f+r}{R-r}\right)$$

ii) the first order approximation of the series expansion for the term relevance (expressions (10) and (13) respectively):

$$w_2(f) = \log\left(\frac{a}{1-a}\right) + \log\frac{I}{R} - \frac{s}{I} + \frac{r}{R} \quad \text{for } f \leq R$$

and

$$w_2(f) = \log\left(\frac{r}{f}\right) + \frac{r}{f} + \log\left(\frac{I}{(1-a)R} - 1\right) \quad \text{for } f > R.$$

iii) the tenth order approximation of the series expansion for the term relevance (expressions (5) and (11) limited to 10 terms for the infinite series):

$$w_3(f) = \log\left(\frac{a}{1-a}\right) + \log\frac{I}{R} - \frac{s}{I} + \sum_{i=1}^{10} \frac{1}{i}\left(\frac{r}{R}\right)^i \quad \text{for } f > R.$$

and

$$w_3(f) = \log\left(\frac{r}{f}\right) + \log\left(\frac{I}{(1-a)R} - 1\right) + \sum_{i=1}^{10} \frac{1}{i}\left(\frac{r}{f}\right)^i \quad \text{for } f > R.$$

iv) the inverse document frequency weight

$$w_4(f) = \log\left(\frac{N}{f}\right) .$$

In each case, the calculated values for r are based on approximation (4); that is, one assumes

$$r = af \quad \text{when } f \leq R$$

and $$r = b + cf \quad \text{when } f > R .$$

The calculated values are shown in Table 2, and the corresponding graphs appear in Figs. 3, 4, and 5 for the three values of the parameter a.

The values of Table 2 indicate that there is practically no difference between the exact relevance formula w_1 and the tenth order approximation w_3. The first order approximation is also fairly good, except in a small neighborhood around f = R. The IDF method, w_4, is acceptable for a = 0.25 and a = 0.53 when f is greater than R. For the case a = 0.75, the IDF values differ from the relevance values. However it can be seen that for f > 30, $w_1 \approx w_4 + 1.20$. This supports the previous claim that the rate of decrease of w_1 is about the same as that of w_4. The development in this section also lends theoretical support to the previously mentioned method by Croft and Harper [11], where the term relevance function was approximated by an inverse document frequency factor plus a constant (expression (3)).

To summarize, if one assumes the conditions given in (4), then it can be seen that the relevance weighting measure first increases nearly linearly. For document frequencies greater than R, the relevance weight decreases at the same rate as the IDF scheme. Furthermore, for the medium frequency terms, the IDF and relevance weights are similar. Since most query terms used in practice may be expected to fall in the medium frequency range where the difference between w(f) and IDF(f) is small, it is not surprising that the available experimental data show little improvement when the simple IDF system is replaced by the term relevance weights.

References

[1] D.H. Kraft and A. Bookstein, Evaluation of Information Retrieval Systems: A Decision Theory Approach, Journal of the ASIS, Vol. 29, 1978, p. 31-34.

[2] S.E. Robertson and K. Sparck Jones, Relevance Weighting of Search Terms, Journal of the ASIS, Vol. 27, No. 3, 1976, p. 129-146.

[3] C.T. Yu, W.S. Luk and M.K. Siu, On Models of Information Retrieval Processes, Information Systems, Vol. 4, No. 3, 1979, p. 205-218.

[4] C.T. Yu and G. Salton, Precision Weighting - An Effective Automatic Indexing Method, Journal of the ACM, Vol. 23, No. 1, 1976, p. 76-88.

[5] K. Sparck Jones, Experiments in Relevance Weighting of Search Terms, Information Processing and Management, Vol. 15, 1979, p. 133-144.

[6] K. Sparck Jones, Search Term Relevance Weighting Given Little Relevance Information, Journal of Documentation, Vol. 35, 1979, p. 30-48.

[7] K. Sparck Jones, Search Term Relevance Weighting - Some Recent Results, Journal of Information Science, Vol. 1, 1980, p. 325-332.

[8] S.E. Robertson, C. J. VanRijsbergen and M.F. Porter, Probabilistic Models of Indexing and Searching, Proc. of ACM-BCS Symposium on Research and Development in Information Retrieval, Cambridge, England, 1980.

[9] G. Salton, A. Wong, and C.T. Yu, Automatic Indexing Using Term Discrimination and Term Precision Measurements, Information Processing and Management, Vol. 12, 1976, p. 43-51.

[10] G. Salton and R.K. Waldstein, Term Relevance Weights in On-Line Information Retrieval, Information Processing and Management, Vol. 14, 1978, p. 29-35.

[11] W. B. Croft and D.J. Harper, Using Probabilistic Models of Document Retrieval Without Relevance Information, Journal of Documentation, Vol. 35, 1979, p. 285-295.

[12] D. J. Harper and C.J. VanRijsbergen, An Evaluation of Feedback in Retrieval Using Co-Occurrence Data, Journal of Documentation, Vol. 34, 1978, p. 189-216.

[13] G. Salton, H. Wu, and C.T. Yu, The Measurement of Term Importance in Automatic Indexing, to be published in Journal of the ASIS.

[14] C.T. Yu, K. Lam, and G. Salton, Optimum Term Weighting in Information Retrieval Using the Term Precision Model, to be published in Journal of the ACM.

[15] K. Sparck Jones, A Statistical Interpretation of Term Specificity and its Application in Retrieval, Journal of Documentation, Vol. 28, 1972, p. 11-21.

Appendix

Lemma 4: For f > R, $\frac{I-s}{R-r} = \frac{1}{(1-a)}\frac{I}{R} - 1$.

Proof: From the original assumptions (4), one has

$$r = af \text{ for } 0 \le f \le R \quad \text{and} \tag{A1}$$
$$r = b+cf \text{ for } R < f \le N \tag{A2}$$

Since r must be continuous at f = R one obtains for r = R

$$aR = b + cR \quad \text{or}$$
$$b = (a-c)R \quad . \tag{A3}$$

Since r = b+cf in the region under consideration in this lemma, one has at f = N

$$R = b + cN \quad . \tag{A4}$$

From (A3) and (A4) one obtains

$$(a-c)R = R-cN \quad .$$

Hence $$C = \frac{(1-a)R}{N-R} \tag{A5}$$

Using (A2) and substituting successively (A3) and (A5) one has

$$\begin{aligned} r &= b + cf \\ &= (a-c)R + cf \\ &= [a - \frac{(1-a)R}{N-R}] R + \frac{(1-a)R}{N-R}] \ f \\ &= [aN-r + \frac{(1-a)f}{N-R}] \ R \quad . \end{aligned}$$

Hence $$\frac{r}{R} = aN-R + \frac{(1-a)f}{N-R} \quad \text{and} \tag{A6}$$

$$1 - \frac{r}{R} = \frac{(1-a)(N-f)}{N-R} \quad . \tag{A7}$$

A similar development is now used to obtain an expression for s/I.

$$\begin{aligned} s &= f-r && \text{by definition} \\ &= f - b - cf && \text{from (A2)} \\ &= f(1-c) - (a-c)\ R && \text{from (A3)} \\ &= f(1 - \frac{(1-a)R}{N-R}) - (a-\frac{(1-a)R}{N-R})\ R \\ &= (N-R-R+aR)f - (aN-R)\frac{R}{N-R} \\ &= \frac{s}{I} = (N-R-R+aR)f - \frac{(aN-R)R}{(N-R)I} && (A8) \end{aligned}$$

Using (A6) and (A8) and noting that N = R+I, it follows that

$$\begin{aligned} \frac{r}{R} - \frac{s}{I} &= \frac{aNI - RI + f - Iaf - Nf + Rf + Rf - aRf + aNR - R^2}{(N-R)I} \\ &= \frac{aNI - RI - Iaf + Rf - aRf + aNr - R^2}{(N-R)I} \\ &= \frac{aNN - RN - Naf + Rf}{(N-R)I} \\ &= \frac{(aN-R)(N-f)}{(N-R)I} && (A9) \end{aligned}$$

The expression needed in the lemma is first rewritten as

$$= \frac{I}{R} \frac{(1-\frac{r}{R}) - (\frac{r}{R} - \frac{s}{I})}{1-\frac{r}{R}} \tag{A10}$$

By substituting (A7) and (A9) into (A10) one obtains:

$$\begin{aligned} \frac{I-s}{R-r} &= \frac{I}{R} \frac{(1-a)N-f)I + (aN-r)(N-f)}{(N-R)I} \cdot \frac{N-R}{(1-a)(N-f)} \\ &= \frac{(1-a)I + (aN-R)}{R(1-a)} \\ &= \frac{I + aR - R}{R(1-a)} && \text{(because } aN - aI = aR) \\ &= \frac{I}{R}\frac{1}{(1-a)} - 1 \quad . && (A11) \end{aligned}$$

This proves the lemma.

Lemma 6:

$$\text{a)} \quad |\log(1 - \frac{r}{f})| < \epsilon/2 \tag{A12}$$

$$\text{b)} \quad |\log(1 - \frac{s}{I})| < \epsilon/2 \quad \text{and} \tag{A13}$$

$$\text{c)} \quad |\log(1 - (1-\frac{r}{R})) - \log(1-\frac{r}{R})| < \epsilon \quad . \tag{A14}$$

Proof: Consider first part a).

Since $\log(1-\frac{R}{f}) \leq \log(1-\frac{r}{f}) < o$ one obtains

$|\log(1-\frac{r}{f})| \leq |\log(1-\frac{R}{f})| < \epsilon/2$

if and only if $\log(1-\frac{R}{f}) > -\epsilon/2$

if and only if $(1-\frac{R}{f}) > e^{-}\epsilon/2$

if and only if $f > R/(1-e^{-}\epsilon/2)$. (A15)

Consider now part (b) of the lemma:

Since

$\log(1-\frac{f}{I}) \leq \log(1-\frac{S}{I}) < o$, one obtains

$|\log(1-\frac{s}{I})| \leq |\log(1-\frac{f}{I})| < \epsilon/2$

if and only if $\log(1-\frac{f}{I}) > -\epsilon/2$

if and only if $(1-\frac{f}{I}) > e^{-}\epsilon/2$

if and only if $f < I(1-e^{-}\epsilon/2)$ (A16)

To prove part (c) it is necessary to use the following identity

$$(x^i - y^i) = (x-y)(x^{i-1} + x^{i-2}y + \ldots + xy^{i-2} + yi-1) \tag{A17}$$

There are two subcases:

i) Assume $r/R \geq 1/2$, so that $(1-N/R) \leq r/R$.

In that case

$$o < \log(1-(1-\tfrac{r}{R})) - \log(1-\tfrac{r}{R}) \qquad \text{(A18)}$$

$$= -\sum_{i=1}^{\infty} \tfrac{1}{i}(1-\tfrac{r}{R})^i + \sum_{i=1}^{\infty} \tfrac{1}{i}(\tfrac{r}{R})^i$$

(using the series of expansion (8))

$$= \sum_{i-1}^{\infty} (2\tfrac{r}{R}-1)\tfrac{1}{i}[(\tfrac{r}{R})^{i-1} + (\tfrac{r}{R})^{i-2}(1-\tfrac{r}{R})$$

$$+\ldots+ \tfrac{r}{R}(1-\tfrac{r}{R})^{i-2}+(1-\tfrac{r}{R})^{i-1}] \text{ (using (A17))}$$

$$= \leq (2\tfrac{r}{R}-1) \sum_{i=1}^{\infty} \tfrac{1}{i} i\,(\tfrac{r}{R})^{i-1} \text{ (since } (1-\tfrac{r}{R}) \leq \tfrac{r}{R})$$

$$= (2\tfrac{r}{R}-1)\frac{1}{1-(r/R)}$$

$$= < (2(\tfrac{1+\epsilon}{2+\epsilon})-1)\,(\frac{1}{1-\frac{1+\epsilon}{2+\epsilon}}) \text{(by the assumption of theorem 5)}$$

$$= \epsilon$$

ii) Assume now $r/R \leq 1/2$, so that $(1 - r/R) \geq r/R$.

In that case

$$0 < -(\log(1-(1-\tfrac{r}{R})) - \log(1-\tfrac{r}{R})) \qquad \text{(A19)}$$

$$= \sum_{i=1}^{\infty} \tfrac{1}{i}(1-\tfrac{r}{R})^i - \sum_{i=1}^{\infty} \tfrac{1}{i}(\tfrac{r}{R})^i \quad \text{(using expression (8))}$$

$$= \sum_{i=1}^{\infty} (1-2\tfrac{r}{R})\tfrac{1}{i}[(1-\tfrac{r}{R})^{i-1} + (1-\tfrac{r}{R})^{i-2}\tfrac{r}{R}$$

$$+\ldots+ (1-\tfrac{r}{R})(\tfrac{r}{R})^{i-2} + (\tfrac{r}{R})^{i-1}]$$

$$\leq (1-2\tfrac{r}{R}) \sum_{i=1}^{\infty} \tfrac{1}{i} i(1-\tfrac{r}{R})^{i-1}$$

$$= (1-2\tfrac{r}{R})\frac{1}{r/R}$$

$$< (1-2\frac{1}{2+\epsilon})\frac{1}{\frac{1}{2+\epsilon}} \qquad \text{(by assumption of theorem 5)}$$

$$= \epsilon$$

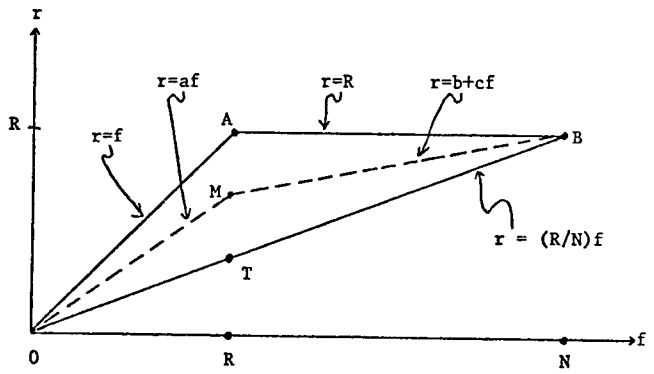
The lemma shows that (A12) holds under conditions (A15). Similarly (A13) is true under the condition (A16). Finally, (A14) is strictly true for both subcases (i) and (ii). Hence the sum of the last four terms of (16) is less than 2 ϵ under the conditions of theorem 5.

QED.

	$\epsilon = 0.5$	$\epsilon = 1$	$\epsilon = 2$
lower frequency bound $R/(1-e^{-\epsilon/2})$	90	51	31
upper frequency bound $I(1-e^{-\epsilon/2})$	221	393	632
$\lvert\log(I/N)\rvert$	0.02	0.02	0.02
$\lvert w(f) - IDF\rvert$	<1.02	<2.02	<4.02

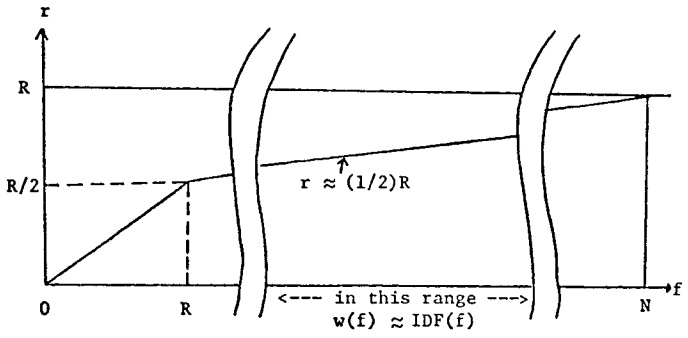
Error Bounds of $|w(f) - IDF|$ of Theorem 5

Table 1



Frequency Spectrum for One Term Relating r and f

Fig. 1



IDF Approximation of Term Relevance

Fig. 2

f	w_1	w_2	w_3	w_4	w_1-w_2	w_1-w_3	w_1-w_4
2	2.837	2.837	2.837	6.234	3.1×10^{-4}	-1.1×10^{-6}	-3.407
5	2.874	2.872	2.874	5.318	2.0×10^{-3}	-7.0×10^{-6}	-2.444
10	2.939	2.931	2.939	4.6250	8.5×10^{-3}	-2.8×10^{-5}	-1.685
15	3.010	2.990	3.010	4.2195	2.0×10^{-2}	-6.3×10^{-5}	-1.209
20	3.085	3.048	3.086	3.9318	3.7×10^{-2}	-1.1×10^{-4}	-0.845
30	2.611	2.594	2.611	3.5263	1.6×10^{-2}	4.1×10^{-10}	-0.916
100	1.468	1.466	1.468	2.3234	2.0×10^{-3}	5.7×10^{-15}	-0.854
200	0.9668	0.9660	0.9668	1.6292	7.6×10^{-4}	-1.5×10^{-16}	-0.6624
300	0.7312	0.7307	0.7312	1.2237	4.8×10^{-4}	-5.5×10^{-17}	-0.4926
400	0.5905	0.5901	0.5905	0.9360	3.6×10^{-4}	2.7×10^{-17}	-0.3456
500	0.4961	0.4958	0.4961	0.7129	3.0×10^{-4}	-1.2×10^{-16}	-0.2168
600	0.4281	0.4279	0.4281	0.5306	2.6×10^{-4}	-9.7×10^{-17}	-0.1024
700	0.3768	0.3765	0.3768	0.3764	2.3×10^{-4}	1.3×10^{-16}	2.8×10^{-3}

a) Comparison of Term Relevance and Inverse Document Frequency
(N = 1020, R = 20, I = 1000, a = 0.25)

f	w_1	w_2	w_3	w_4	w_1-w_2	w_1-w_3	w_1-w_4
2	4.085	4.084	4.085	6.234	$1.4\ 10^{-3}$	$-4.4\ 10^{-7}$	-2.149
5	4.172	4.162	4.172	5.318	$9.6\ 10^{-3}$	$-2.7\ 10^{-6}$	-1.146
10	4.335	4.292	4.335	4.625	$4.2\ 10^{-2}$	$-1.1\ 10^{-5}$	-0.290
15	4.532	4.422	4.532	4.219	$1.1\ 10^{-1}$	$-1.9\ 10^{-5}$	0.313
20	4.778	4.553	4.778	3.931	$2.2\ 10^{-1}$	$1.2\ 10^{-4}$	0.847
30	4.066	3.982	4.067	3.526	$8.4\ 10^{-2}$	$1.5\ 10^{-6}$	0.540
100	2.602	2.595	2.602	2.323	$6.9\ 10^{-3}$	$4.0\ 10^{-12}$	0.279
200	1.931	1.929	1.931	1.629	$1.9\ 10^{-3}$	$5.3\ 10^{-15}$	0.302
300	1.582	1.581	1.582	1.223	$1.0\ 10^{-3}$	$8.8\ 10^{-16}$	0.359
400	1.353	1.353	1.353	0.936	$6.4\ 10^{-4}$	$8.8\ 10^{-16}$	0.417
500	1.189	1.189	1.189	0.712	$4.7\ 10^{-4}$	$8.8\ 10^{-16}$	0.477
600	1.064	1.063	1.064	0.530	$3.6\ 10^{-4}$	$1.0\ 10^{-20}$	0.534
700	0.9638	0.964	0.9638	0.3764	$2.9\ 10^{-4}$	$-6.6\ 10^{-16}$	0.587

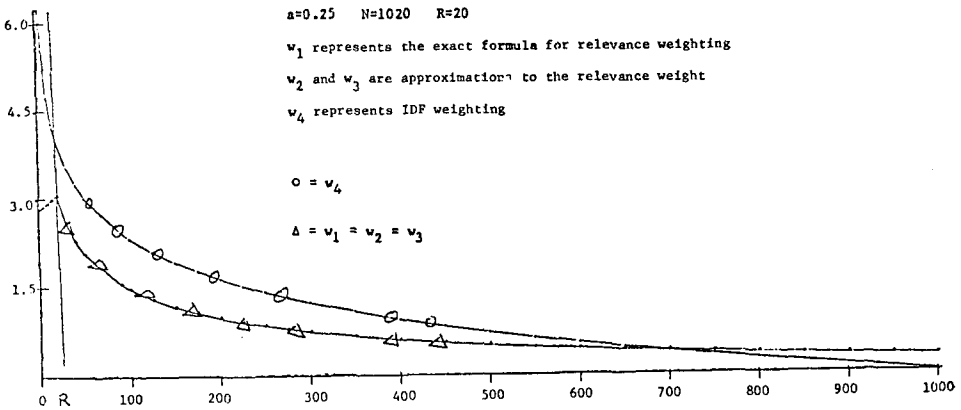
b) Comparison of Term Relevance and Inverse Document Frequency
(N = 1020, R = 20, I = 1000, a = 0.53)

f	w_1	w_2	w_3	w_4	w_1-w_2	w_1-w_3	w_1-w_4
2	5.088	5.085	5.088	6.234	2.9×10^{-3}	-1.2×10^{-7}	-1.15
5	5.217	5.196	5.217	5.318	2.0×10^{-2}	-7.8×10^{-7}	-0.101
10	5.478	5.383	5.478	4.625	9.5×10^{-2}	-2.6×10^{-7}	0.853
15	5.833	5.569	5.833	4.219	2.6×10^{-1}	3.3×10^{-4}	1.614
20	6.392	5.755	6.379	3.931	6.3×10^{-1}	1.2×10^{-2}	2.461
30	5.300	5.105	5.300	3.526	1.9×10^{-1}	8.5×10^{-5}	1.774
100	3.590	3.576	3.589	2.323	1.3×10^{-2}	1.2×10^{-10}	1.267
200	2.844	2.840	2.844	1.629	3.3×10^{-3}	1.3×10^{-15}	1.215
300	2.443	2.441	2.443	1.223	1.5×10^{-3}	1.3×10^{-15}	1.22
400	2.172	2.171	2.172	0.936	9.1×10^{-4}	6.6×10^{-16}	1.236
500	1.970	1.969	1.970	0.712	6.1×10^{-4}	4.4×10^{-16}	1.258
600	1.811	1.811	1.811	0.530	4.5×10^{-4}	2.2×10^{-16}	1.281
700	1.681	1.681	1.681	0.3764	3.5×10^{-4}	-2.2×10^{-16}	1.304

c) Comparion of Term Relevance and Inverse Document Frequency
(N = 1020, R = 20, I = 1000, a = 0.75)

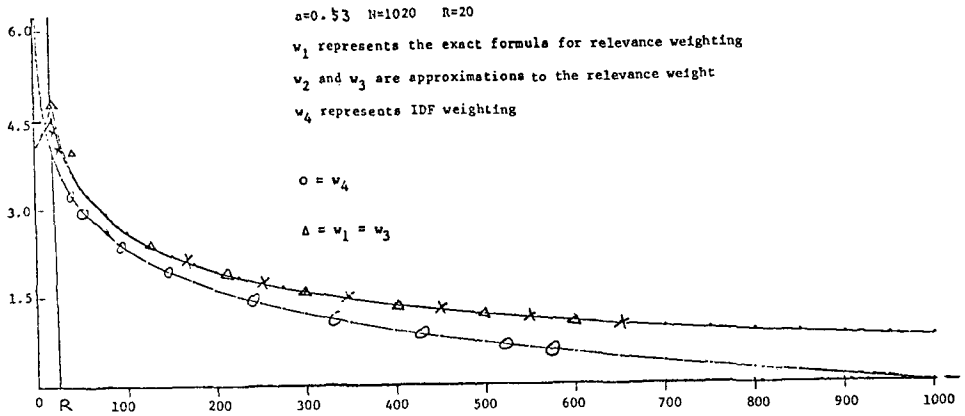
Experimental Comparison of Term Relevance and IDF
for Three Values of a (0.25, 0.53, 0.75)

Table 2



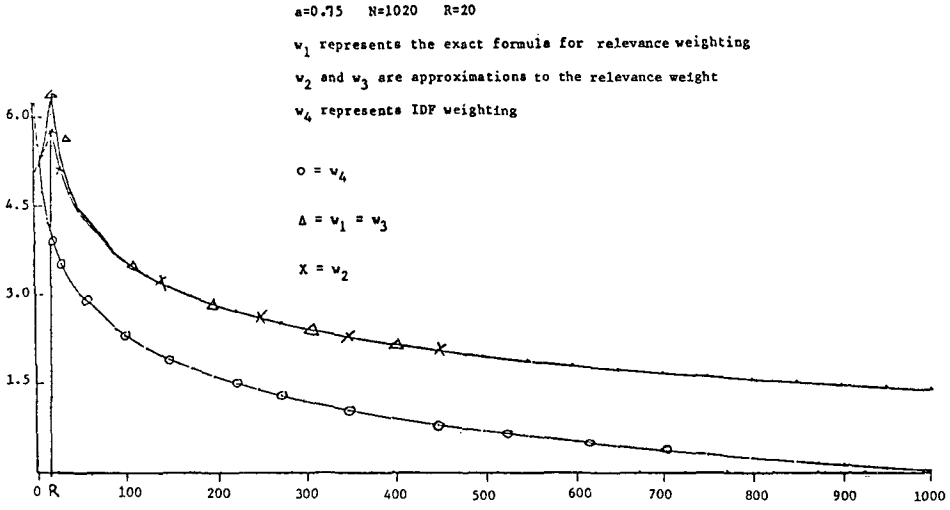
Graphical Output for Data of Table 2(a)

Figure 3



Graphical Output for Data of Table 2(b)

Figure 4



Graphical Output for Data of Table 2(c)

Figure 5