Term Relevance Weights in On-Line Information Retrieval

by

G. Salton and R.K. Waldstein

TR 77-316

Computer Science Department
Cornell University
Ithaca, NY 14853

TERM RELEVANCE WEIGHTS IN ON-LINE INFORMATION RETRIEVAL

G. Salton and R.K. Waldstein*

ABSTRACT

Considerable evidence exists to show that the use of term relevance weights is beneficial in interactive information retrieval. Various term weighting systems are reviewed. An experiment is then described in which information retrieval users are asked to rank query terms in decreasing order of presumed importance prior to actual search and retrieval. The experimental design is examined, and various relevance ranking systems are evaluated, including fully automatic systems based on inverse document frequency parameters, human rankings performed by the user population, and combinations of the two.

Dept. of Computer Science, Cornell University, Ithaca, NY 14853

1. ON LINE INFORMATION RETRIEVAL

Substantial changes have taken place over the last few years in the manner in which bibliographic searches are being conducted. In former times, a search would take place off-line without contact between the user population and the stored information files. As a result, the search activity would often be a "hit or miss" operation, in the sense that the retrieval results might be very good or very poor, depending on the particular way in which the search requirements could be formulated and the search would be handled.

More recently, interactive retrieval strategies have been developed in which continuous contact is maintained during a search between the users and the automated information store. In such an environment tentative query formulations can be submitted using console entry devices, and the query statements and search results can be adjusted repeatedly until the users are satisfied with the retrieval results. Many practical bibliographic retrieval systems operate currently in such an "on-line" mode, and texts and manuals exist describing their detailed operations. [1]

Two principal types of user-system interaction can be distinguished in on-line retrieval environments, termed "presearch" and "postsearch" interaction, respectively. [2] As the name indicates, presearch interaction takes place prior to the actual search operation and consists in using vocabulary and thesaurus displays to refine the query statements. Postsearch interaction, on the

other hand, utilizes information about previously retrieved items--for example, document titles or abstracts--to reformulate the search requests. In such a case, an attempt is often made to render the query statements as similar as possible to the descriptions of items judged to be useful, or, on the contrary as different as possible from item descriptions known to be extraneous. In this fashion, a reformulated query is likely to retrieve additional items that the user may be expected to accept, and/or fewer items that are likely to be rejected.

One of the problems in interactive retrieval is the substantial effort required on the part of the user population in the reformulation of the search requests. In many instances acceptable results are obtained only for experienced users that are familiar with the properties of the search vocabulary and with the collection environment.

To alleviate the burden placed on the users in the query reformulation process, various automatic or semi-automatic feedback methods have been developed including the use of "objective" term weighting systems, and of document relevance assessments supplied by the users in the course of the search operations. Thus, term weights may be obtained automatically by using specified characteristics of the query formulations and document descriptions in a given collection. For example, terms occurring in documents judged to be relevant to the users' information needs may receive higher weights than those occurring in documents identified as not relevant. This leads to the so-called "rele-

vance feedback" process implemented in various experimental and operational retrieval situations. [3,4]

In the remainder of this study, a variety of simple automatic or semi-automatic term alteration methods are described which are easy to implement while also being effective in interactive retrieval.

2. TERM WEIGHTING SYSTEMS

A great many term weighting systems have been described in the literature, whose function it is to differentiate among the elements of the document or query descriptions based on their presumed order of importance. For present purposes it may suffice to cite the well-known term frequency (TF) criterion in which the term weight depends on the frequency of occurrence of a given term in each individual document; the inverse document frequency (IDF) system in which the TF factor is divided by the number of distinct documents to which a term is assigned; and finally the discrimination value (DV) strategy which is based on the ability of a given term to distinguish the various documents of a collection from each other. [5,6,7] These term weighting systems utilize objective criteria to compute the various indicators of term importance. Thus, the inverse document frequency system is based on the assumption that terms indicative of subject content occur frequently in certain items of a collection, but that their overall frequency of occurrence in the whole collection is rather small. Hence a term weight is chosen which is directly proportional to the frequency of occurrence of the term

in each individual document, and inversely proportional to the number of documents in a collection to which the term is assigned.

A possible formula for w_i^k, the weight of term k in document i, which has been found to be effective in retrieval is

$$w_i^k = (F_i^k) \cdot ([\log_2 N] - [\log_2 B^k] + 1) \qquad (1)$$

The first factor in the formula is the frequency of occurrence of term k in document i, whereas the second factor is the inverse document frequency (IDF) factor where N represents the total number of documents in the collection, and B^k the number of documents containing term k. [5] For simplicity, equation (1) may then be rewritten as

$$w_i^k = (F_i^k) \cdot (IDF^k) \qquad (1')$$

With the development of interactive retrieval environments, new term weighting systems suggest themselves which depend at least in part on subjective information obtained from the user during the retrieval process. Thus, the term weights might depend on specific information about individual term values obtained from the users, or on document relevance data. In the latter case, an attempt is made to increase the weights of terms that occur generally in the documents identified as relevant to the search requests, while at the same time downgrading the weights of terms that are found mostly in the nonrelevant items.

Term weighting systems based on the use of relevance information are examined in the remainder of this study.

3. TERM RELEVANCE COMPUTATIONS

A variety of term weighting systems based on user relevance information have been proposed in the literature, under the general heading of term accuracy, term precision, or term relevance systems. [8,9,10] The term accuracy model is based on the assumption that the document frequency is inversely related to the probability of occurrence of the terms in the relevant documents of a collection (that is, the smaller the number of documents to which a term is assigned, the greater is its occurrence probability in the relevant items). Based on this assumption one can show that a thesaurus of low-frequency terms used to add semantically related terms to document and query descriptions will prove effective in retrieval. [9] The same is true of a phrase assignment process which assigns narrow (low frequency) phrases to replace broader (higher frequency) terms.

An equivalent, but formally more satisfactory process consists in making assumptions about term occurrences and relevance characteristics within a document collection to derive an optimal term weighting function capable of ranking the terms in decreasing order of presumed usefulness in retrieval. Consider, in particular, the following two assumptions:

a) the occurrence distribution of the terms is assumed to be independent in the relevant documents within a collection, and also in the nonrelevant documents of a collection, and

b) the probable relevance of a document with respect to a

query is assumed to be based not only on the presence of the search terms in the document description but also on their absence from the description.

The latter assumption is noncontroversial in the sense that it is not counterintuitive to base a query-document similarity computation on both the matching as well as the nonmatching terms. The term independence assumption is more hazardous--obviously two cooccurring terms that are assigned to a single document only exhibit perfect dependence (rather than independence). However for many medium- and high-frequency terms, the independence assumption is in fact not unreasonable. [11]

These assumptions lead to a formally defined term ranking function which may be expressed as the logarithm of the ratio of the proportion of relevant documents to which a term is assigned to the proportion of nonrelevant items containing that term. [8] More precisely, p^k, the term precision of term k is defined as

$$p^k = \log \left\{ \frac{r^k}{R - r^k} \Big/ \frac{n^k - r^k}{N - n^k - R + r^k} \right\} \qquad (2)$$

where r^k is the number of relevant documents with respect to some query containing term k,

n^k is the total number of nonrelevant documents containing term k,

N is the total number of documents in the collection

and R is the total number of relevant documents in the collection with respect to the query.

It should be noted that unlike the inverse document frequency

weighting function of equation (1) which is usable directly for term weighting purposes, the term precision expression of equation (2) is not document-dependent. An actual term weight may be computed by using, for example, the product of the term frequency and the term precision as follows:

$$w_i^k = (F_i^k) \cdot (P^k). \qquad (3)$$

It may be shown formally that a term weighting function proportional to P^k is necessarily effective in retrieval; that is, for any given recall point (for any proportion of relevant items retrieved), the retrieval precision (the number of nonrelevant items that are rejected) obtained with the P^k factor is at least as large as the retrieval precision produced by binary-weighted terms. [10]

Various problems arise in connection with the use of weighting functions based on the occurrence probabilities in the relevant and nonrelevant documents of a collection. The most immediate difficulty is the nonavailability of the required relevance assessments of documents with respect to queries. For practical purposes, it then becomes necessary to use a set of test queries for which full query-document relevance assessments are in fact obtainable. The occurrence characteristics of the terms in the test sample are then used to compute the P^k factors for all the terms present in the query and/or document sets. These term relevance factors can then be used in two different ways:

a) retrospectively, by applying the corresponding weights (as in equation (3)) to the same test queries and docu-

ments from which the weights were originally derived;

b) predictively, by applying the weights derived from the test collection to new queries not previously utilized.

The retrospective use is of course unobjectionable and may be expected to produce an upper bound for the retrieval performance for the given query and document test sets. On the other hand, a retrospective use of term weighting functions is impractical, since it is hardly useful to perform a new retrieval operation after the full relevance information has first been obtained for all documents with respect to all queries.

The predictive use of the term precision weights is, however, theoretically useful. In practice, the process is difficult to implement for two main reasons:

a) the queries and documents not originally included in the test sets may be expected to contain new terms for which term precision information is then unavailable;

b) the relative importance of the terms may be expected to vary from user to user, and hence from query to query; in these circumstances the application of term precision information derived from one query-document collection to a different environment becomes questionable. [8]

One concludes that the predictive use of term precision weights may fail in cases where the user population is not homogeneous, or the test collection lacks comprehensiveness.

This fact is illustrated in the output of Table 1, where a

collection of 425 documents from Time magazine is used with 83 user queries. The test set used to derive the term precision weights consisted of 41 queries out of the 83 that were available. The document terms are weighted in each case using the standard term frequency (TF) factors; the query terms, on the other hand, use either the term frequency alone, or the function incorporating the term precision (equation (3)), retrospectively as well as predictively. In each case, the retrieval precision values, averaged over the 41 or 42 user queries are given for certain fixed values of the recall. It may be seen that large improvements in retrieval precision are obtained for the retrospective case, but a slight loss in performance actually results when the precision weights are used predictively for a new query set. The main reason is the small size of the experimental test set: the 41 test queries contained 229 distinct terms of which only 85 were also included in the 42 queries used for the predictive experiment. When precision weights can be generated for only a small portion of the existing terms, a successful outcome is not likely.

One concludes from the foregoing discussion that term relevance information is theoretically useful, but that in practice its utilization raises operational problems that are difficult to overcome. A more direct use of term relevance properties is described in the next section.

4. USER RELEVANCE RANKING

In view of the problems inherent in the utilization of automatically produced relevance weights, one may ask whether term relevance measures could be supplied manually. In particular, the users might be shown a list of pertinent terms--for example, terms included in the original query formulation, or terms related to the query terms extracted from a thesaurus or dictionary. Each user would then be asked to rank the terms for his query in decreasing order of presumed importance, 1 indicating the most important term, 2 the next most important, and so on down to rank m for the least important term. These results can then be converted into term weights in such a way that the terms of lowest rank receive the highest weight. The term weights themselves are usable as before for the query-document matching process.*

A manual term ranking system such as the one previously described and illustrated in Fig. 1 was implemented for the 41 test queries in conjunction with the 425 Time magazine articles in world affairs. The output of Table 2 contains the recall-precision output, averaged over the 41 test queries. The first column, representing the standard term frequency weighting, is added for control purposes. Subsequent columns of Table 2 correspond to the following weighting systems:

---

* One particular rank-weight conversion system assigns term weights ranging from 1 to 2, 1 being the weight assigned to the term of highest rank m. [10]

a) column 2 represents the inverse document frequency weighting scheme (equations (1) or (1')) which was previously shown to produce a high standard of performance [5,6];

b) column 3 is the precision weight system used retrospectively (equation (3)); this output may represent the optimum term weighting system attainable for the test collection;

c) columns 4 and 5 correspond to the manual term ranking system where R^k represents the relevance weight assigned to term k as produced by the rank-weight conversion system; in both cases, the inverse document frequency weight system is utilized for the document terms; the query terms use the relevance weights with term frequency alone (column 4), and with inverse document frequency included (column 5).

The output of Table 2 shows that the manual reranking system by itself is not as powerful as the automatic inverse document frequency systems (column 4 versus column 2). However, when the manual reranking system is used in addition to the inverse document frequency (column 5), substantial additional improvements are obtained. Indeed for low recall, the resulting hybrid system is nearly as powerful as the optimum retrospective precision weight scheme of column 3; high performance standards are also produced at the high recall end.

The conclusion appears inescapable that the on-line facilities

available in modern retrieval environments are usable to good advantage by asking the customers to judge the presumed importance of displayed query or document terms. This raises the question of the importance of user training in an interactive retrieval environment. In particular, one would expect that informed users familiar with retrieval system operations would be able to produce more effective term ranks than uninformed users.

Such a conclusion can indeed be confirmed by comparing the term rankings produced by seven uninformed users participating in the experiment with those generated by four informed users. [12] In both cases, the users were graduate students in computer science, the latter group consisting of persons with training in information retrieval. The rankings produced for one particular query by the 11 different users are shown in Table 3. Table 3 also contains a comparison with the automatic term precision weights for the corresponding terms. It may be seen that the informed users were able correctly to recognize the importance of the geographic terms "Kenya", "Tanganyika", and "Uganda", whereas the uninformed users preferred the less important term "independence", and the much less useful "federation".

In conclusion one can say that term relevance information appears useful as a component of term weighting systems used in retrieval. Since the predictive use of the automatic term precision values is unreliable except when large test collections are available for computational purposes, a simple manual term relevance ranking system might be used in interactive environments.

The term ranking systems may be expected to become especially useful when the relevance ranks are supplied by informed user populations.

REFERENCES

1. P.W. Lancaster and S. Fayen, Information Retrieval On-Line, John Wiley and Sons, New York, 1973.

2. M.E. Lesk and G. Salton, Interactive Search and Retrieval Methods Using Automatic Information Displays, in The SMART Retrieval System, G. Salton, editor, Prentice Hall Inc., Englewood Cliffs, N.J., 1971, Chapter 25.

3. G. Salton, Relevance Feedback and the Optimization of Retrieval Effectiveness, in the SMART Retrieval System, G. Salton, editor, Prentice Hall Inc., Englewood Cliffs, N.J. 1971, Chapter 15.

4. C.O Vernimb and G. Steven. ENDS--European Nuclear Documentation Service, Nuclear Engineering and Design, Vol. 25, 1973, p. 325-333.

5. K. Sparck Jones, A Statistical Interpretation of Term Simplicity and its Application to Retrieval, Journal of Documentation, Vol. 28, No. 1, March 1972, p. 11-20.

6. G. Salton and C.S. Yang, On the Specification of Term Values in Automatic Indexing, Journal of Documentation, Vol. 29, No. 4, December 1973, p. 351-372.

7. G. Salton, Dynamic Information and Library Processing, Prentice Hall Inc., Englewood Cliffs, N.J. 1975.

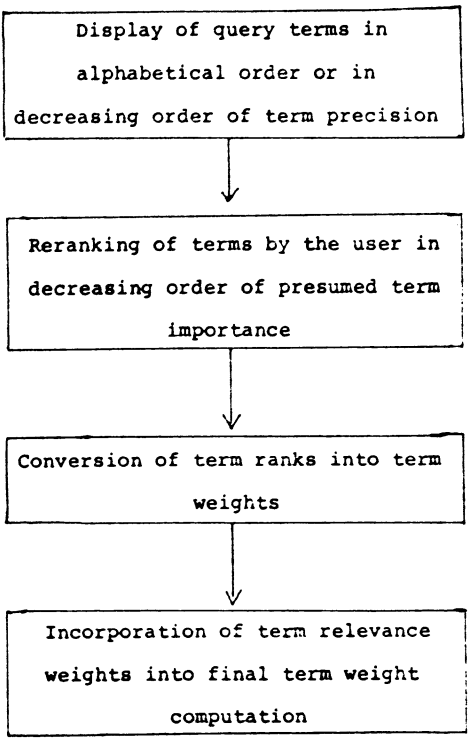
8. S.E. Robertson and K. Sparck Jones, Relevance Weighting of Search Terms, Journal of the ASIS, Vol. 27, No. 3, May-June 1976, p. 129-146.

9. C.T. Yu and G. Salton, Effective Information Retrieval Using Term Accuracy, ACM Communications, Vol. 20, No. 3, March 1971, p. 135-142.

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

11. C.T. Yu, G. Salton, and M. K. Siu, Effective Automatic Indexing Using Term Addition and Deletion, to appear in Journal of the ACM.

12. R.H. Waldstein, Weighting Methodologies for Queries and Documents, Master's Thesis, Computer Science Department, May 1977.



Simplified Term Relevance Ranking System

Fig. 1.

Recall	Term Frequency Documents; Term Frequency Queries	Term Frequency Documents; Retrospective Relevance Weights for Queries		Term Frequency Documents; Predictive Relevance Weights for Queries	
0	.4163	.4391	+5%	.3989	-4%
.1	.4159	.4391	+6	.3987	-4
.2	.4159	.4391	+6	.3963	-5
.3	.4115	.4363	+6	.3954	-4
.4	.4040	.4310	+7	.3934	-3
.5	.3936	.4262	+8	.3836	-3
.6	.3495	.3910	+12	.3411	-2
.7	.3292	.3681	+12	.3185	-3
.8	.3204	.3601	+12	.3111	-3
.9	.3044	.3476	+14	.2942	-3
1.0	.3011	.3420	+14	.2920	-3

Predictive and Retrospective Query Relevance Weights

(Time Collection: 425 documents, 41 retrospective user queries, 42 predictive user queries, using word stemming and term frequency weights) (adapted from [12])

Table 1

Recall	Term Frequency Documents; Term Frequency Queries	TF * IDF Documents; TF · IDF Queries		R Pre
	$w_i^k = F_i^k$	$w_i^k = F_i^k \cdot IDF^k$		w_i^k
0	.4163	.4440	+7%	
.1	.4159	.4440	+7	
.2	.4159	.4440	+7	
.3	.4115	.4418	+7	
.4	.4040	.4341	+7	
.5	.3936	.4277	+9	
.6	.3495	.3971	+14	
.7	.3292	.3703	+12	
.8	.3204	.3641	+14	
.9	.3044	.3500	+15	
1.0	.3011	.3428	+14	

Automatic and Manually App

(Time Collection: 425 do
word stemming, term freq
ment frequency

Tab

TF * IDF Documents; etrospective cision Weights for Queries $= F_i^k \cdot P^k$		TF * IDF Documents; Informed User Rankings for Queries $w_i^k = F_i^k \cdot R^k$		TF * IDF Documents; IDF Queries * Informed User Rankings $w_i^k = F_i^k \cdot IDF^k \cdot R^k$	
.4590	+10%	.4228	+2%	.4517	+9%
.4590	+10	.4228	+2	.4517	+9
.4590	+10	.4223	+2	.4517	+9
.4569	+11	.4201	+2	.4479	+9
.4527	+12	.4127	+2	.4399	+9
.4460	+13	.4054	+2	.4293	+9
.4073	+17	.3757	+7	.3966	+13
.3869	+17	.3545	+8	.3766	+14
.3778	+23	.3460	+8	.3713	+16
.3612	+19	.3381	+11	.3577	+18
.3530	+24	.3309	+10	.3507	+16

lied Term Relevance Weights

cuments, 41 user queries;
uency (TF), inverse docu-
(IDF) weighting)

le 2

QUERY: Federation of East Africa to be formed by Kenya, Tanganyika, and Uganda when Kenya gains its independence from Britain on December 12.

Query Terms	Automatic Precision Weights Pk (Term Rank)		Informed User Ranks A	B	C	D	Uninformed User Ranks A	B	C	D	E	F	G
Federation	10	(6)	7	6	5	3	8	4	1	3	3	2	3
Britain	16	(5)	6	7	4	3	6	5	6	5	7	8	11
Formed	0	(11)	8	8	9	3	10	6	10	8	11	9	10
Africa	4	(9)	9	9	7	3	10	4	12	2	2	1	2
Independence	54	(4)	5	4	6	3	1	1	2	1	1	3	1
Gains	2	(10)	10	10	10	3	7	7	8	4	8	4	4
December	7	(7)	11	11	11	3	9	7	11	7	10	6	8
East	6	(8)	2	5	8	2	5	3	7	6	9	7	9
Kenya	68	(3)	1	1	1	1	2	2	3	8	4	5	5
Tanganyika	139	(2)	3	2	2	1	3	2	4	8	5	9	6
Uganda	844	(1)	4	3	3	1	4	2	5	8	6	9	7

Informed Versus Uninformed User Rankings
for One Test Query
(adapted from [12])

Table 3