Term Weighting Approaches in Automatic Text Retrieval

Gerard Salton*
Chris Buckley

87-881
November 1987

Department of Computer Science
Cornell University
Ithaca, New York 14853-7501

*This study was supported in part by the National Science Foundation under grants IST 83-16166 and IRI 87-02735.

Term Weighting Approaches in Automatic Text Retrieval

Gerard Salton and Chris Buckley *

Abstract

The experimental evidence accumulated over the past 20 years indicates that text indexing systems based on the assignment of appropriately weighted single terms produce retrieval results that are superior to those obtainable with other more elaborate text representations. These results depend crucially on the choice of effective term weighting systems. This paper summarizes the insights gained in automatic term weighting, and provides baseline single term indexing models with which other more elaborate content analysis procedures can be compared.

*Department of Computer Science, Cornell University, Ithaca, NY 14853.
This study was supported in part by the National Science Foundation under grants IST 83-16166 and IRI 87-02735.

1. Automatic Text Analysis

In the late 1950s, Luhn first suggested that automatic text retrieval systems could be designed based on a comparison of content identifiers attached both to the stored texts and to the users' information queries. [1] Typically, certain words extracted from the texts of documents and queries would be used for content identification; alternatively, the content representations could be chosen manually by trained indexers familiar with the subject areas under consideration and with the contents of the document collections. In either case, the documents would be represented by *term vectors* of the form

$$D = (t_i, t_j, ..., t_p) \tag{1}$$

where each t_k identifies a content term assigned to some sample document D. Analogously, the information requests, or queries, would be represented either in vector form, or in the form of Boolean statements. Thus, a typical query Q might be formulated as

$$Q = (q_a, q_b, ..., q_r) \tag{2}$$

$$or \qquad Q = (q_a \ and \ q_b) \ or \ (q_c \ and \ q_d \ and \ ...) \ or \ ... \tag{3}$$

where q_k once again represents a term assigned to query Q.

A more formal representation of the term vectors of Expressions (1) and (2) is obtained by including in each vector all possible content terms allowed in the system, and adding term weight assignments to provide distinctions among the terms. Thus, if w_{dk} (or w_{qk}) represents the weight of term t_k in document D (or query Q), and t terms in all are available for content representation, the term vectors for document D and query Q can be written as

$$D = (t_0, w_{d_0}; \ t_1, w_{d_1}; \ ...; \ t_t, w_{dt})$$

$$\text{and} \qquad Q = (q_0, w_{q_0}; \ q_1, w_{q_1}; \ ...; \ q_t, w_{qt}) \tag{4}$$

In the foregoing formulation, the assumption is that w_{dk} (or w_{qk}) is equal to 0 when term k is not assigned to document D (or query Q), and that w_{dk} (or w_{qk}) equals 1 for the assigned terms.

Given the vector representations of Expression (4), a *query-document similarity* value may be obtained by comparing the corresponding vectors, using for example the conventional vector pro-

duct formula

$$similarity(Q,D) = \sum_{k=1}^{t} w_{qk} \cdot w_{dk} \tag{5}$$

When the term weights are restricted to 0 and 1 as previously suggested, the vector product of Expression (5) measures the number of terms that are jointly assigned to query Q and document D.

In practice, it has proven useful to provide a greater degree of discrimination among terms assigned for content representation than is possible with weights of 0 and 1 alone. In particular, term weights in decreasing term importance order could be assigned, in which case the weights w_{dk} (or w_{qk}) could be allowed to vary continuously between 0 and 1, the higher weight assignments near 1 being used for the most important terms, whereas lower weights near 0 would characterize the less important terms. In some circumstances, it may also be useful to use normalized weight assignments, where the individual term weights depend to some extent on the weights of other terms in the same vector. A typical term weight using a vector length normalization factor is $\frac{w_{dk}}{\sqrt{\sum_{vector} (w_{di})^2}}$ for documents (or $\frac{w_{qk}}{\sqrt{\sum_{vector} (w_{qi})^2}}$ for queries). When a length normalized term weighting system is used with the vector similarity function of expression (5), one obtains the well-known cosine vector similarity formula that has been used extensively with the experimental Smart retrieval system [2,3]:

$$similarity(Q,D) = \frac{\sum_{k=1}^{t} w_{qk} \cdot w_{dk}}{\sqrt{\sum_{k=1}^{t} (w_{qk})^2 \cdot \sum_{k=1}^{t} (\tilde{w}_{dk})^2}} \tag{6}$$

A vector matching system performing global comparisons between query and document vectors provides ranked retrieval output in decreasing order of the computed similarities between Q and D. Such a ranked output is useful because controls are now available over the size of the retrieved document set, and iterative retrieval strategies based on successive query reformulations are simplified. A system that first retrieves those items thought to be of main interest to the users will necessarily prove helpful in interactive information retrieval.

In designing automatic text retrieval systems, two main questions must be faced: the first is the choice of appropriate content units to be included in the document and query representations, and the second is the determination of the term weights capable of distinguishing the important terms from those less crucial for content identification.

Concerning first the choice of content terms, various possibilities must be considered. In most of the early experiments, *single terms* alone were used for content representation, often consisting of words extracted from the texts of documents and from natural language query formulations. [3 - 7] In many cases, quite effective retrieval output has been obtained using single term content representations. Ultimately, however, sets of single terms cannot provide complete identifications of document content. For this reason, many enhancements in content analysis and text indexing procedures have been proposed over the years in an effort to generate complex text representations. The following possibilities have been considered in this connection:

a) The generation of sets of *related terms* based on the statistical co-occurrence characteristics of the words in certain contexts within the document collection. The assumption normally made is that words that co-occur with sufficient frequency in the documents of a collection are in fact related to each other. [8 - 11]

b) The formation of *term phrases* consisting of one or more governing terms (the phrase heads) together with corresponding dependent terms (the phrase components). Phrases are often chosen by using word frequency counts and other statistical methods, possibly supplemented by syntactic procedures designed to detect syntactic relationships between governing and dependent phrase components. [12 - 17]

c) The use of word grouping methods of the kind provided by *thesauruses*, where classes of related words are grouped under common headings; these class headings can then be assigned for content identification instead of the individual terms contained in the classes. [18 - 20] Alternatively, term relationships useful for content identification may also be obtainable by using existing machine-readable dictionaries and lexicons. [21 - 24]

d) The construction of *knowledge bases* and related artificial intelligence structures designed to

represent the content of the subject area under consideration; entries from the knowledge base are then used to represent the content of documents and queries. [25 - 30]

From the beginning, it was evident that the construction and identification of complex text representations was inordinately difficult. In particular, it became clear that most automatically derived term dependencies were valid only locally in the documents from which the dependent term groups were originally extracted; this implies that dependent term groups could not be counted upon to produce useful content identifiers in new document contexts different from those originally used. [11] The experiences gained with the use of automatically generated term phrases proved similarly discouraging: for some collections, improvements in retrieval effectiveness of up to 20 per cent (in search recall and precision) were obtainable by using phrase identifiers instead of single terms; but for other collections these same phrase procedures did not furnish any improvements at all. Moreover, even sophisticated syntactic analysis programs could not be relied upon to produce useful complex content identifiers. [16]

As for the use of pre-constructed vocabulary schedules and term classifications, the problem is that viable procedures for the construction of effective vocabulary tools covering subject areas of reasonable scope appear to be completely lacking. The same goes for the construction of knowledge bases designed to reflect the structure of discourse areas. Until more becomes known about the desired form and content of dictionaries and thesauruses, little gain should be expected from these tools in text analysis and document indexing.

In reviewing the extensive literature accumulated during the past 25 years in the area of retrieval system evaluation, the overwhelming evidence is that the judicious use of single term identifiers is preferable to the incorporation of more complex entities extracted from the texts themselves, or obtained from available vocabulary schedules. [31-37] Two main problems appear in producing complex text identifiers:

a) When stringent conditions are used for the construction of complex identifiers, typified by the use of restrictive frequency criteria and limited co-occurrence contexts for the recognition of term phrases, then few new identifiers are likely to become available, and the performance of

the retrieval system with complex identifiers will differ only marginally from the results obtainable with single term indexing.

b) On the other hand, when the construction criteria for the complex entities are relaxed, then some good identifiers are obtained, but also many marginal ones that do not prove useful. Overall, the single term indexing will generally be preferred.

When single terms are used for content identification, distinctions must be introduced between individual terms, based on their presumed value as document descriptors. This leads to the use of term weights attached to the item identifiers. The considerations controlling the generation of effective weighting factors are outlined briefly in the next section.

2. Term Weight Specification

The main function of a term weighting system is the enhancement of retrieval effectiveness. Effective retrieval depends on two main factors: on the one hand, items likely to be relevant to the user's needs must be retrieved; on the other hand, items likely to be extraneous must be rejected. Two measures are normally used to assess the ability of a system to retrieve the relevant and reject the nonrelevant items of a collection, known as *recall* and *precision*, respectively. Recall is the proportion of relevant items retrieved, measured by the ratio of the number of relevant retrieved items to the total number of relevant items in the collection; precision, on the other hand, is the proportion of retrieved items that are relevant, measured by the ratio of the number of relevant retrieved items to the total number of retrieved items.

In principle, a system is preferred that produces both high recall by retrieving everything that is relevant, and also high precision by rejecting all items that are extraneous. The recall function of retrieval appears to be best served by using broad, high-frequency terms that occur in many documents of the collection. Such terms may be expected to pull out many documents, including many of the relevant documents. The precision factor, however, may be best served by using narrow, highly specific terms that are capable of isolating the few relevant items from the mass of nonrelevant ones. In practice, compromises are normally made by using terms that are

broad enough to achieve a reasonable recall level without at the same time producing unreasonably low precision.

The differing recall and precision requirements favor the use of composite term weighting factors that contain both recall- and precision-enhancing components. Three main considerations appear important in this connection:

a) Terms that are frequently mentioned in individual documents, or document excerpts, appear to be useful as recall-enhancing devices. This suggests that a *term frequency* (tf) factor be used as part of the term weighting system measuring the frequency of occurrence of the terms in the document or query texts. Term frequency weights have been used for many years in automatic indexing environments. [1 - 4]

b) Term frequency factors alone cannot insure acceptable retrieval performance. Specifically, when the high frequency terms are not concentrated in a few particular documents, but instead are prevalent in the whole collection, all documents tend to be retrieved, and this affects the search precision. Hence a new collection-dependent factor must be introduced which favors terms concentrated in a few documents of a collection. The well-known *inverse document frequence* (idf) (or inverse collection frequency) factor performs this function. The idf factor varies inversely with the number of documents n to which a term is assigned in a collection of N documents. A typical idf factor may be computed as $\log N/n$. [38]

Term discrimination considerations suggest that the best terms for document content identification are those able to distinguish certain individual documents from the remainder of the collection. This implies that the best terms should have high term frequencies but low overall collection frequencies. A reasonable measure of term importance may then be obtained by using the product of the term frequency and the inverse document frequency (tf $\times$ idf). [39 - 41]

The term discrimination model has been criticized because it does not exhibit well substantiated theoretical properties. This is in contrast with the probabilistic model of information retrieval where the relevance properties of the documents are taken into account, and a theoretically valid *term relevance* weight is derived. [42 - 44] The term relevance weight, defined as the

proportion of relevant documents in which a term occurs divided by the proportion of nonrelevant items in which the term occurs is, however, not immediately computable without knowledge of the occurrence properties of the terms in the relevant and nonrelevant parts of the document collection. A number of methods have been proposed for estimating the term relevance factor in the absence of complete relevance information, and these have shown that under well-defined conditions the term relevance can be reduced to an inverse document frequency factor of the form $\log((N-n)/n)$. [45-46] The composite (tf $\times$ idf) term weighting system is thus directly relatable to other theoretically attractive retrieval models.

A third term weighting factor, in addition to the term frequency and the inverse document frequency, appears useful in systems with widely varying vector lengths. In many situations, short documents tend to be represented by short term vectors, whereas much larger term sets are assigned to the longer documents. When a large number of terms are used for document representation, the chance of term matches between queries and documents is high, and hence the larger documents have a better chance of being retrieved than the short ones. Normally, all relevant documents should be treated as equally important for retrieval purposes. This suggests that a *normalization factor* be incorporated into the term weighting formula to equalize the length of the document vectors. Assuming that w represents the weight of term t, the final term weight might then be defined as $w / \sum_{vector\ i} w_i$, or $w / \sqrt{\sum_{vector\ i} (w_i)^2}$.

In the preceding discussion of term weighting systems both documents and queries were assumed to be represented by sets, or vectors, of weighted terms. Term weighting systems have also been applied to Boolean query statements, and extended Boolean systems have been devised in which Boolean query statements are effectively reduced to vector form. [47-54] The previous considerations regarding term weighting thus apply to some extent also to Boolean query processing.

3. Term Weighting Experiments

A number of term weighting experiments are described in the remainder of this note in which combinations of term frequency, collection frequency, and length normalization components are

used with 6 document collections of varying size, covering different subject areas. In each case, collections of user queries are used for retrieval purposes and the performance is averaged over the number of available user queries. For each experiment the average search precision is computed for 3 different recall points, including a low recall of 0.25, an average recall of 0.50, and a high recall of 0.75. This average search precision is then further averaged for all available user queries. In addition, to the precision measure, the rank of the weighting methods in decreasing performance order is used as an evaluation criterion. A total of 1800 different combinations of term weight assignments were used experimentally, of which 287 were found to be distinct. A rank of 1 thus designates the best performance, and 287 the worst.

In the present experiments each term weight combination is described by using two triples, representing respectively the term frequency, collection frequency, and vector normalization factors for document terms (first triple), and query terms (second triple). The principal weighting components are defined in Table 1. Three different term frequency components are used, including a binary weight (b), the normal term frequency (t), and a normalized term frequency (n) which lies between 0.5 and 1.0. The three collection frequency components represent multipliers of 1(x) that disregards the collection frequency, a conventional inverse collection frequency factor (f), and a probabilistic inverse collection frequency (p). Finally the length normalization factor may be absent (x as the third component) or present (c). (In the previously mentioned full set of 1800 different term weight assignments, additional weighting components not included in Table 1 were also tried. These additional components did not supply any fundamentally new insights or advantages.)

Table 2 shows actual formulas for some well-known term weighting systems. The *coordination-level* match which simply reflects the number of matching terms present in documents and queries, respectively, is described by the sextuple bxx·bxx. Similarly, the probabilistic binary term independence system which uses binary document terms, but a probabilistic inverse collection frequency weight for the query terms, is represented as bxx·bpx. A typical complex term weighting scheme, described as tfc·nfx, uses a normalized tf × idf weight for document terms, and an enhanced, but unnormalized tf × idf factor for the queries. (Since the query vectors remain con-

stant for all documents of a collection, a query normalization simply multiplies all query-document similarity measurements by a constant factor which leaves the final document ranking unaffected.)

The six collections used experimentally are characterized by the statistics of Table 3. The smallest collection is a biomedical (MED) collection, consisting of 1033 documents and 30 queries, whereas the largest collection (INSPEC) comprises 12684 documents and 84 queries, covering the computer engineering areas. In all cases, the query vectors are much shorter than the corresponding document vectors.

The NPL (National Physical Laboratory) collection of 11429 documents and 100 queries was available in indexed form only (that is, in the form of document and query vectors) and not in original natural language form. This may explain its somewhat peculiar make-up. Both the document and the query vector are much shorter in the NPL collection than in the other collections, and the variation in query length (2.36 for a mean number of 7.16 query terms) is very small. Furthermore, the term frequencies are especially low for the NPL collection: each query term appears precisely once in a query, and the average frequency of the terms in the documents is only 1.21. In these circumstances, the term frequency weighting and length normalization operations cannot perform their intended function. One may conjecture that the NPL index terms are carefully chosen, and may in fact represent specially controlled terms rather than freely chosen natural language entries.

Typical evaluation output is shown in Tables 4(a), and 4(b). With a few minor exceptions, the results for the 5 collections of Table 4(a) are homogeneous, in the sense that the best results are produced by the same term weighting systems for all collections, and the same holds also for the poorest results. The results of Table 4(a) do however differ substantially from those obtained for the NPL collection in Table 4(b). Considering first the results of Table 4(a), the following conclusions are evident:

a) Methods 1 and 2 produce comparable performances for all collections, the length normalization is important for the documents, and the enhanced query weighting is effective for the queries. These methods are recommended for conventional natural language texts and text

abstracts.

b) Method 3 does not include the normalization operation for vector length, nor the enhanced query weights. This unnormalized (tf × idf) weighting method is poor for collections such as CRAN and MED where very short query vectors are used with little deviation in the query length. In such cases, enhanced query weights (n factor) prove important.

c) Method 4 represents the best of the probabilistic weighting systems. This method is less effective than the enhanced weighting schemes of methods 1 and 2. It fails especially for collections such as CISI and INSPEC where long query vectors are used, and the term discrimination afforded by query term weighting is essential.

d) Methods 5 to 7 represent, respectively, the classical inverse document frequency weighting, the probabilistic binary term independence system, and the classical term frequency weighting. As can be seen, these methods are generally inferior for all collections.

e) The coordination level matching of binary vectors represents one of the worst possible retrieval strategies.

The results of Table 4(b) for the NPL collection differs markedly from those of Table 4(a). Here the probabilistic schemes using binary query weights and unnormalized document vectors are preferred. This is a direct result of the special nature of the queries and documents for that collection: the very short queries with little length deviation require fully weighted query terms (b = 1), and the normally effective term frequency weights should be avoided because many important terms will then be downgraded in the short document vectors. An enhanced term frequency weight (n factor), or a full weight (b = 1) is therefore preferred. Retrieval results obtained for NPL were used earlier to claim superiority for the probabilistic term weighting system. [55] The results of Table 4 do not support this contention for conventional natural language documents and queries.

4. Recommendations

The following conclusions may be drawn from the experimental evidence reported in this study:

Query vectors:

a) term frequency component

- for short query vectors, each term is important; enhanced query term weights are thus preferred: first component n

- long query vectors require a greater discrimination among query terms based on term occurrence frequencies: first component t

- the term frequency factor can be disregarded when all query terms have occurrence frequencies equal to 1.

b) collection frequency component

- inverse collection frequency factor f is very similar to the probabilistic term independence factor p; best methods use f.

c) normalization component

- query normalization does not affect query-document ranking or overall performance; use x.

Document vectors:

a) term frequency component

- for technical vocabulary and meaningful terms (CRAN, MED collections), use enhanced frequency weights: first component n.

- for more varied vocabulary, distinguish terms by conventional frequency weights: first component t

- for short document vectors possibly based on controlled vocabulary, use fully weighted terms: first component b = 1.

b) collection frequency component

- inverse document frequency factor f is similar to probabilistic term independence weight p: normally use f.

- for dynamic collections with many changes in the document collection make-up, the f factor requires updating; in that case disregard second component: use x.

c) length normalization component

- when the deviation in vector lengths is large, as it normally is in text indexing systems, use length normalization factor c.

- for short document vectors of homogeneous length, the normalization factor may be disregarded; in that case use x.

The following single term weighting systems should be used as a standard for comparison with enhanced text analysis systems using thesauruses and other knowledge tools to produce complex multi-term content identifications:

best document weighting	:	tfc, nfc (or tpc, npc)
best query weighting	:	nfx, tfx, bfx (or npx, tpx, bpx).

References

[1] H.P. Luhn, A Statistical Approach to the Mechanized Encoding and Searching of Literary Information, IBM Journal of Research and Development, 1:4, October 1957, 309-317.

[2] G. Salton, editor, The Smart Retrieval System--Experiments in Automatic Document Retrieval, Prentice Hall Inc., Englewood Cliffs, NJ, 1971.

[3] G. Salton and M.J. McGill, Introduction to Modern Information Retrieval, McGraw Hill Book Co., New York, 1983.

[4] C.J. van Rijsbergen, Information Retrieval, Butterworths, London, Second Edition, 1979.

[5] H.P. Luhn, A New Method of Recording and Searching Information, American Documentation, 4:1, 1955, 14-16.

[6] M. Taube and I.S. Wachtel, The Logical Structure of Coordinate Indexing, American Documentation, 3:4, 1952, 213-218.

[7] J.W. Perry, Information Analysis for Machine Searching, American Documentation, 1:3, 1950, 133-139.

[8] C.J. van Rijsbergen, A Theoretical Basis for the Use of Cooccurrence Data in Information Retrieval, Journal of Documentation, 33:2, June 1977, 106-119.

[9] G. Salton, C. Buckley, and C.T. Yu, An Evaluation of Term Dependence Models in Information Retrieval, Lecture Notes in Computer Science, G. Salton and H.J. Schneider, editors, 146, Springer-Verlag, Berlin, 1983, 151-173.

[10] C.T. Yu, C. Buckley, K. Lam and G. Salton, A Generalized Term Dependence Model in Information Retrieval, Information Technology: Research and Development, 2:4, October 1983, 129-154.

[11] M.E. Lesk, Word-Word Associations in Document Retrieval Systems, American Documentation, 20:1, January 1969, 27-38.

[12] P.H. Klingbiel, Machine Aided Indexing of Technical Literature, Information Storage and Retrieval, 9:2, February 1973, 79-84.

[13] P.H. Klingbiel, A Technique for Machine Aided Indexing, Information Storage and Retrieval, 9:9, September 1973, 477-494.

[14] M. Dillon and A. Gray, Fully Automatic Syntax-Based Indexing, Journal of the ASIS, 34:2, March 1983, 99-108.

[15] K. Sparck Jones and J.I. Tait, Automatic Search Term Variant Generation, Journal of Documentation, 40:1, March 1984, 50-66.

[16] J.L. Fagan, Experiments in Automatic Phrase Indexing for Document Retrieval: A Comparison of Syntactic and Non-Syntactic Methods, Doctoral Thesis, Report 87-868, Department of Computer Science, Cornell University, Ithaca, NY, September 1987.

[17] A.F. Smeaton, Incorporating Syntactic Information into a Document Retrieval Strategy: An Investigation, Proc. 1986 ACM-SIGIR Conference on Research and Development in Information Retrieval, Pisa, Italy, Association for Computing Machinery, New York, 1986, 103-113.

[18] K. Sparck Jones, Automatic Keyword Classification for Information Retrieval, Butterworths, London, 1971.

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

[20] R.T. Dattola, Experiments with Fast Algorithms for Automatic Classification, in The Smart Retrieval System--Experiments in Automatic Document Processing, G. Salton, editor, Chapter 12, Prentice Hall Inc., Englewood Cliffs, NJ, 1971, 265-297.

[21] D.E. Walker, Knowledge Resource Tools for Analyzing Large Text Files, in Machine

Translation: Theoretical and Methodological Issues, Sergei Nirenburg, editor, Cambridge University Press, Cambridge, England, 1987, 247-261.

[22] H. Kucera, Uses of On-Line Lexicons, Proceedings First Conference of the U.W. Centre for the New Oxford English Dictionary: Information in Data, University of Waterloo, 1985, 7-10.

[23] R.A. Amsler, Machine-Readable Dictionaries, Annual Review of Information Science and Technology, M.E. Williams, editor, Vol. 19, Knowledge Industry Publication Inc., White Plains, NY, 1984, 161-209.

[24] E.A. Fox, Lexical Relations: Enhancing Effectiveness of Information Retrieval Systems, ACM SIGIR Forum, 15:3, 1980, 5-36.

[25] W.B. Croft, User-Specified Domain Knowledge for Document Retrieval, Proceedings of the 1986 ACM Conference on Research and Development in Information Retrieval, Pisa, Italy, Association for Computing Machinery, New York, 1986, 201-206.

[26] R.H. Thompson and W.B. Croft, An Expert System for Document Retrieval, Proceedings of the Expert Systems in Government Symposium, IEEE Computer Society Press, Washington, DC, 1985, 448-456.

[27] W.B. Croft, Approaches to Intelligent Information Retrieval, Information Processing and Management, 23:4, 1987, 249-254.

[28] K. Sparck Jones, Intelligent Retrieval, in Intelligent Information Retrieval: Proceedings of Informatics, Vol. 7, Aslib, London, 1983, 136-142.

[29] E.A. Fox, Development of the Coder System: A Testbed for Artificial Intelligence Methods in Information Retrieval, Information Processing and Management, 23:4, 1987, 341-366.

[30] G. Salton, On the Use of Knowledge Based Processing in Automatic Text Retrieval, Proceedings of 49th Annual Meeting of the ASIS, Learned Information, Medford, NJ, 1986, 277-287.

[31] D.R. Swanson, Searching Natural Language Text by Computer, Science, 132:3434, October 1960, 1099-1104.

[32] C.W. Cleverdon and E.M. Keen, Aslib-Cranfield Research Project, Vol. 2, Test Results, Cranfield Institute of Technology, Cranfield, England, 1966.

[33] C.W. Cleverdon, A Computer Evaluation of Searching by Controlled Languages and Natural Language in an Experimental NASA Database, Report ESA 1/432, European Space Agency, Frascati, Italy, July 1977.

[34] F.W. Lancaster, Evaluation of the Medlars Demand Search Service, National Library of Medicine, Bethesda, MD, January 1968.

[35] D.C. Blair and M.E. Maron, An Evaluation of Retrieval Effectiveness for a Full-Text Document Retrieval System, Communications of the ACM, 28:3, March 1985, 289-299.

[36] G. Salton, Another Look at Automatic Text Retrieval Systems, Communications of the ACM, 29:7, July 1986, 648-656.

[37] G. Salton, Recent Studies in Automatic Text Analysis and Document Retrieval, Journal of the ACM, 20:2, April 1973, 258-278.

[38] K. Sparck Jones, A Statistical Interpretation of Term Specificity and its Application in Retrieval, Journal of Documentation, 28:1, March 1972, 11-21.

[39] G. Salton and C.S. Yang, On the Specification of Term Values in Automatic Indexing, Journal of Documentation, 29:4, December 1973, 351-372.

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

[41] G. Salton, C.S. Yang and C.T. Yu, A Theory of Term Importance in Automatic Text Analysis, Journal of the ASIS, 26:1, January-February 1975, 33-44.

[42] A. Bookstein and D.R. Swanson, A Decision Theoretic Foundation for Indexing, Journal of the ASIS, 26:1, January-February 1975, 45-50.

[43] W.S. Cooper and M.E. Maron, Foundation of Probabilistic and Utility Theoretic Indexing, Journal of the ACM, 25:1, 1978, 67-80.

[44] S.E. Robertson and K. Sparck Jones, Relevance Weighting of Search Terms, Journal of the ASIS, 27:3, 1976, 129-146.

[45] W.B. Croft and D.J. Harper, Using Probabilistic Models of Information Retrieval without Relevance Information, Journal of Documentation, 35:4, December 1975, 285-295.

[46] H. Wu and G. Salton, A Comparison of Search Term Weighting: Term Relevance versus Inverse Document Frequency, ACM SIGIR Forum, 16:1, Summer 1981, 30-39.

[47] T. Noreault, M. Koll and M.J. McGill, Automatic Ranked Output from Boolean Searches in SIRE, Journal of the ASIS, 27:6, November 1977, 333-339.

[48] T. Radecki, Incorporation of Relevance Feedback into Boolean Retrieval Systems, Lecture Notes in Computer Science, 146, G. Salton and H.J. Schneider, editors, Springer Verlag, Berlin, 1982, 133-150.

[49] C.D. Paice, Soft Evaluation of Boolean Search Queries in Information Retrieval Systems, Information Technology: Research and Development, 3:1, 1983, 33-41.

[50] S.C. Cater and D.H. Kraft, A Topological Information Retrieval System (TIRS) Satisfying the Requirements of the Waller-Kraft Wish List, Proceedings of Tenth Annual ACM-SIGIR Conference on Research and Development in Information Retrieval, C.T. Yu and C.J. van Rijsbergen, editors, Association for Computing Machinery, New York, 1987, 171-180.

[51] S.K.M. Wong, W. Ziarko, V.V. Raghavan, and P.C.N. Wong, Extended Boolean Query Processing in the Generalized Vector Space Model, Report, Department of Computer Science, University of Regina, Regina, Canada, 1986.

[52] S.K.M. Wong, W. Ziarko and P.C.N. Wong, Generalized Vector Space Model in Information Retrieval, Proceedings of Eighth Annual ACM-SIGIR Conference on Research and Development in Information Retrieval, Association for Computing Machinery, New York, 1985, 18-25.

[53] H. Wu, On Query Formulation in Information Retrieval, Doctoral Dissertation, Cornell University, January 1981.

[54] G. Salton, E.A. Fox and H. Wu, Extended Boolean Information Retrieval, Communications of the ACM, 26:11, November 1983, 1022-1036.

[55] W.B. Croft, A Comparison of the Cosine Correlation and the Modified Probabilistic Model, Information Technology: Research and Development, 3:2, April 1984, 113-114.

Term Frequency Component		
b	1.0	binary weight equal to 1 for terms present in a vector (term frequency is ignored)
t	tf	raw term frequency (number of times a term occurs in a document or query text)
n	$0.5+0.5\ \frac{tf}{\max\ tf}$	augmented normalized term frequency (tf factor normalized by maximum tf in the vector, and further normalized to lie between 0.5 and 1.0)
Collection Frequency Component		
x	1.0	no change in weight; use original term frequency component (b, t or n)
f	$\log \frac{N}{n}$	multiply original tf factor by an inverse collection frequency factor (N is total number of documents in collection, and n is number of documents to which a term is assigned)
p	$\log \frac{N-n}{n}$	multiply tf factor by a probabilistic inverse collection frequency factor
Normalization Component		
x	1.0	no change; use factors derived from term frequency and collection frequency only (no normalization)
c	$1/\sqrt{\sum_{vector} w_i^2}$	use cosine normalization where each term weight w is divided by a factor representing Euclidian vector length

Table 1. Term Weighting Components

Weighting System	Document Term Weight	Query Term Weight
Best fully weighted system tfc·nfx	$\frac{tf \cdot \log \frac{N}{n}}{\sqrt{\sum_{vector} \left[tf_i \cdot \log \frac{N}{n_i} \right]^2}}$	$\left[0.5 + \frac{0.5\ tf}{\max tf} \right] \cdot \log \frac{N}{n}$
Best weighted probabilistic weight nxx·bpx	$0.5 + \frac{0.5\ tf}{\max tf}$	$\log \frac{N-n}{n}$
Classical idf weight bfx·bfx	$\log \frac{N}{n}$	$\log \frac{N}{n}$
Binary term independence bxx·bpx	1	$\log \frac{N-n}{n}$
Standard tf weight: txc·txx	$\frac{tf}{\sqrt{\sum_{vector} (tf_i)^2}}$	tf
Coordination level bxx·bxx	1	1

Table 2. Typical Term Weighting Formulas

Collection	Number of Vectors (Documents or Queries)	Average Vector Length (Number of Terms)	Standard Deviation of Vector Length	Average Frequency of Terms in Vectors	Percentage of Terms in Vectors with Frequency 1
CACM					
documents	3204	24.52	21.21	1.35	80.93
queries	64	10.80	6.43	1.15	88.68
CISI					
documents	1460	46.55	19.38	1.37	80.27
queries	112	28.29	19.49	1.38	78.36
CRAN					
documents	1398	53.13	22.53	1.58	69.50
queries	225	9.17	3.19	1.04	95.69
INSPEC					
documents	12684	32.50	14.27	1.78	61.06
queries	84	15.63	8.66	1.24	83.78
MED					
documents	1033	51.60	22.78	1.54	72.70
queries	30	10.10	6.03	1.12	90.76
NPL					
documents	11429	19.96	10.84	1.21	84.03
queries	100	7.16	2.36	1.00	100.00

Table 3. Collection Statistics (including average vector length and standard deviation of vector lengths)

	Term Weighting Methods	Rank of Method and Ave. Precision	CACM 3204 docs 64 queries	CISI 1460 docs 112 queries	CRAN 1397 docs 225 queries	INSPEC 12684 docs 84 queries	MED 1033 docs 30 queries	Aver for 5 Colle
1.	Best fully weighted (tfc·nfx)	Rank P	1 0.3630	14 0.2189	19 0.3841	3 0.2626	19 0.5628	11.2
2.	Weighted with inverse frequency f not used for docs (txc·nfx)	Rank P	25 0.3252	14 0.2189	7 0.3950	4 0.2626	32 0.5542	16.4
3.	Classical tf × idf No normalization (tfx·tfx)	Rank P	29 0.3248	22 0.2166	219 0.2991	45 0.2365	132 0.5177	84.4
4.	Best weighted probabilistic (nxx·bpx)	Rank P	55 0.3090	208 0.1441	11 0.3899	97 0.2093	60 0.5449	86.2
5.	Classical idf without normalization (bfx·bfx)	Rank P	143 0.2535	247 0.1410	183 0.3184	160 0.1781	178 0.5062	182
6.	Binary independence probabilistic (bxx·bpx)	Rank P	166 0.2376	262 0.1233	154 0.3266	195 0.1563	147 0.5116	159
7.	Standard of weights cosine normalization (original Smart) (txc·txx)	Rank P	178 0.2102	173 0.1539	137 0.3408	187 0.1620	246 0.4641	184
8.	Coordination level binary vectors (bxx·bxx)	Rank P	196 0.1848	284 0.1033	280 0.2414	258 0.0944	281 0.4132	260

Table 4(a). Performance Results for 8 Term Weighting Methods Averaged over 5 Collections

Evaluation	Best fully weighted tfc·nfx	Weighted restricted f txc·nfx	Classical tf × idf tfx·tfx	Best probabilistic nxx·bpx	Classical idf system bxx·bpx	Binary independence txc·txx	Standard weight bxx·bxx	Coordination level
Rank	116	62	149	2	23	8	172	83
Average Precision	0.1933	0.2170	0.1846	0.2752	0.2406	0.2596	0.1750	

Table 4(b). Performance Results for NPL Collection (11429 docs, 100 queries)