A Blueprint for Automatic Indexing

G. Salton*
Department of Computer Science
Cornell University
Ithaca, New York 14853

Abstract:

This note summarizes some of the currently available insights in automatic indexing. The emphasis is on aspects that are expected to be useful in a practical automatic indexing applications. The discussion is necessarily cursory, but the references will lead interested readers to a deeper treatment of the indexing problem.

1. The Indexing Environment

In information retrieval, the stored documents and records are normally identified by sets of terms or keywords that are collectively used to represent the document content. The task of assigning the terms to the individual documents is known as indexing. The indexing task is obviously crucial for retrieval because failures in the indexing policy immediately lead to retrieval failures. Indeed if the indexing is insufficiently exhaustive--that is, if the chosen index terms do not properly reflect all the subject areas covered by a given document--it may not be possible to retrieve a document when it is needed. On the other hand, when the assigned terms are too broad and insufficiently specific, it may not be possible to reject a document that is clearly extraneous. Retrieval performance is often measured by the ability of the system to retrieve the items wanted by the users (the recall factor) and at the same time to reject the extraneous items that are not wanted (the precision factor). A highly exhaustive indexing which uses reasonably specific terms for document content representation is believed to lead to high recall as well as high precision.

At the present time two principal indexing strategies are used in operational retrieval environments:

a) In most situations, the indexing is performed manually by trained indexers, or subject experts, who assign to each document terms that may be freely chosen or may be taken from a controlled list of acceptable terms. Typically between 5 and 15 distinct terms are then assigned to each item for content representation.

b) In some system, an automatic so-called full text indexing system is used where all the words (except for a few common function words) included in a document, or document excerpt, are collectively assigned as index terms.

The quality of the manual indexing is dependent on the experience and background of the individual indexers. The full text indexing system is not subject to the same variability; however the assumption that each text word is equally important for content representation is subject to question.

In the current practice, the indexing task is assumed to be document-specific. That is, each document is treated as a self-contained entity, and the terms assigned to that document do not depend on terms assigned to any other document in the collection. In fact, the usefulness of the index terms varies greatly depending on the collection environment. The term "computer" might not be acceptable as a content indicator for a computer science collection because that term would have to be assigned to all items in the collection; on the other hand, the term "computer" might be essential for documents on medical computer applications included in a collection of medical documents.

During the past twenty years, refined automatic indexing techniques have been developed that use more discrimination than the previously mentioned full text indexing systems and produce high-quality indexing assignments. The available evidence shows that these automatic indexing systems will outperform both of the currently used methods based on human indexing and on full text indexing. The main features of these automatic systems are outlined in the remainder of this note.

2. Term Importance and Term Frequency

The early attempts at automatic indexing were based on the automatic manipulation of machine-readable texts, or text excerpts. Typically, the occurrence frequencies of the individual text words were obtained for the documents of a collection, and these frequency counts were then converted into indications of term usefulness. At first, the role of the frequency parameters was not well understood.

Two basic rules were however accepted as important from the beginning: [1,2]

Rule 1: A relationship exists between the occurrence frequency of a term in a document, or document excerpt, and the importance of that term for the content representation of that document.

Rule 2: A relation exists between the occurrence frequency of a term in a collection of documents and its importance for content representation.

Considering first Rule 1, it is generally the case that when a word occurs many times in a document text, that word represents an important concept in that document. On the other hand, when a word occurs many times in all the documents of a collection its assignment as a content indicator will not distinguish the documents from each other. Hence if the document frequency of a term is defined as the total occurrence frequency in a given document, while the collection frequency is the number of documents in a collection in which the term occurs, the best terms for purposes of content identification will have a high document frequency in individual documents but a low overall collection frequency.

Given a sample collection of documents, or document excerpts in a given subject area, it is now possible to introduce a frequency-based weighting function IDF_{ij}, reflecting the presumed importance of term T_j for the content representation of document D_i. This function, known as the inverse document frequency function is defined as

$$IDF_{ij} = \frac{\text{Document Frequency of } T_j \text{ in } D_i}{\text{Collection Frequency of } T_j}$$

The IDF function is easy to compute for the terms of a document collection and provides the basis for a high-quality automatic indexing strategy using as index terms words extracted from document excerpts or abstracts. The basic five-step process is outlined in Table 1.

1. Identify the individual words occurring in the documents of a collection.

2. Use a stop list of common function words (and, of, or, but, the, etc.) to delete from the texts the high frequency function words that are insufficiently specific for purposes of content representation.

3. Use an automatic suffix stripping routine to reduce each remaining word to word stem form; this reduces to a common form all words exhibiting the same stem (for example, analysis, analyzer, analyzing, etc., are all reduced to stem ANALY).

4. For each remaining word stem compute the IDF weighting function.

5. Represent each document D_i by the set of word stems together with the corresponding IDF weights; that is,

$$D_i = (T_1, IDF_{i1}; T_2, IDF_{i2}; \ldots; T_t, IDF_{it}).$$

Simple Automatic Indexing Strategy Based on Term Extraction

Table 1

No explanation has been offered so far for the second and third steps in Table 1. The stop list eliminates from consideration the terms of highest frequency; the word stem process, on the other hand, increases the assignment frequency of the terms to the documents by replacing certain lower-frequency specific terms by the corresponding higher-frequency more general word stems. Thus the two procedures of steps two and three of Table 1 produce an apparently contradictory effect. This phenomenon is examined in more detail in the next section.

3. Term Specificity and Term Discrimination

It is often believed that an indexing process such as the one in Table 1 cannot be adequate because no recognizable linguistic procedures are included. At the very least synonyms might be recognized by using a thesaurus, and word phrases rather than single words could be used for purposes of content representation.

When thesauruses and term phrases are used for indexing purposes, certain relationships are recognized between the terms used for content representation. To understand the function of these constructs it is useful to return to the notion of term specificity. Consider first the indexing characteristics of very specific and very broad terms:

a) Very specific terms may be expected to be assigned to very few documents. Such terms are effective in rejecting marginal documents but some relevant documents may not be retrievable when the terms are too precise; hence specific terms favor precision at the expense of recall.

b) Very broad terms may be assigned to very many documents. They are effective in retrieving the relevant documents, but a broad term will also reach nonrelevant items that are extraneous. Hence broad terms favor recall at the expense of precision.

Since in retrieval, the recall as well as the precision are important--one wants to retrieve a reasonable proportion of relevant items without at the same time capturing too many nonrelevant ones--it is apparent that the terms used for indexing should exhibit an appropriate level of specificity. In particular, the terms should not be too broad, nor too narrow. The question of term specificity can be approached by studying what happens when a particular term is assigned as an index term to the documents of a collection.

The basic purpose of the indexing task generally, and of an individual term assignment in particular, is the generation of clustering properties among the documents in such a way that related items will be identified by similar term sets, while unrelated documents will receive distinct identifiers. Assuming that document similarity is represented graphically as an (inverse) function of document distance--the more similar the documents, the closer they appear in the graph--the assignment of terms to documents produces the following effect:

a) When very broad terms are assigned as content identifiers, the distances between documents become very small because the broad terms will be assigned to many documents in a collection thus rendering the documents similar to each other. The document "space" is then very dense as shown in the graph of Fig. 1(a).

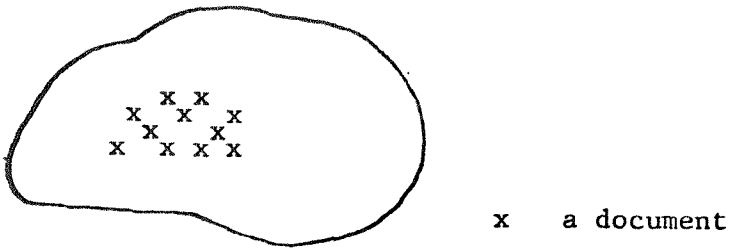

Fig. 1(a) Compressed Document Space Produced
by Assignment of Broad Terms

b) When very specific terms are assigned as content identifiers, the relative distances between the documents of a collection remain unchanged, because specific terms are rarely assigned and almost all documents are unaffected. An originally unclustered document space therefore remains unclustered as shown in Fig. 1(b).

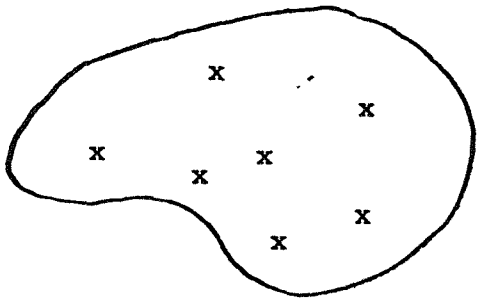

Fig. 1(b) Unclustered Document Space Produced
by Assignment of Very Specific Terms

c) Medium frequency terms that are assigned neither too often nor too rarely produce a clustering effect in that they distinguish the documents to which they are assigned from the remainder. Assuming that the term assignment is correct, documents which are to be jointly retrieved will appear close to each other in the document space, and removed from the other documents which are to be rejected, as shown in the example of Fig. 1(c).

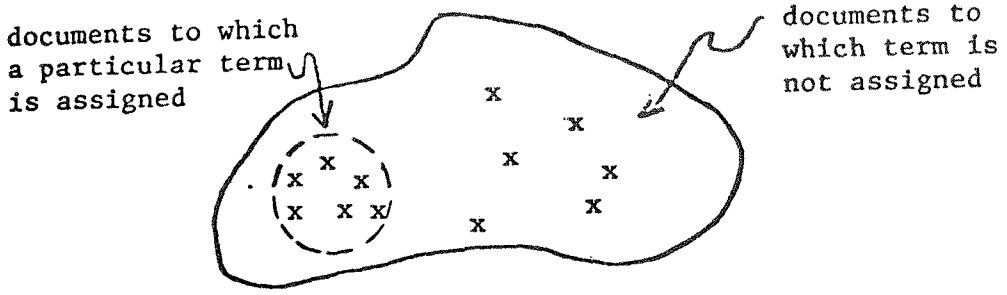

Fig. 1(c) Clustering Effect Produced by Assignment of Term of Correct Specificity

The clustering effect of each individual term can be measured by comparing the space density before a term is assigned with the density after the term is assigned. The discrimination value of a term can then be measured as the difference between the two densities. In particular

$$DV_j = Q - Q_j$$

where DV_j represents the discrimination value of term j, and Q and Q_j are the space densities before and after the assignment of term j, respectively. The space density can be measured as the average pair-wise similarity between all document pairs in a collection. [3,4] Assuming that the situation of Fig. 1 accurately reflects the term assignment effect, the following term discrimination values are obtained:

a) broad terms exhibit negative discrimination values because the density after the term assignment will be greater than before;

b) specific terms exhibit discrimination values close to 0, because their assignment does not alter the space density;

c) medium frequency terms exhibit positive discrimination values because their assignment distinguishes a small class of items from the remainder.

By analogy with the inverse document frequency function, a discrimination value weighting function $DISC_{ij}$ can be defined for each term T_j occurring in document D_i as

$$DISC_{ij} = (\text{Document Frequency of } T_j \text{ in } D_i) * (DV_j).$$

The currently available evidence indicates that the discrimination value weighting function is as effective in retrieval as the previously mentioned inverse document frequency function.

4. Use of Term Relationships

The term discrimination analysis will clarify the role of thesaurus and term phrases in automatic indexing. A thesaurus is a term grouping device which assembles into common classes certain terms believed to be synonymous or semantically related. Instead of assigning a particular term to a given document, the thesaurus makes it possible to assign a complete thesaurus class. A thesaurus class has a broader scope than each individual term included in the class. Thus when a thesaurus is used to group terms which by themselves are too specific, terms of near-zero discrimination value can be transformed into terms with positive discrimination values. The use of a thesaurus corresponds to a left-to-right transformation on the frequency spectrum of Fig. 2.

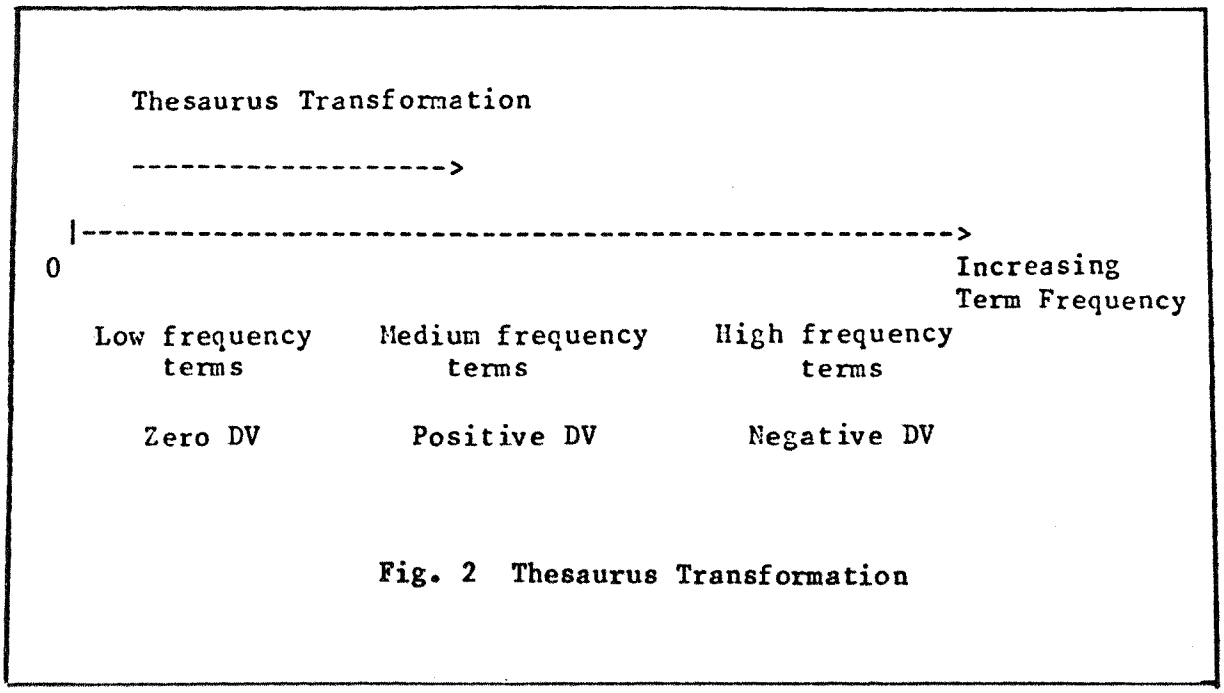

Fig. 2 Thesaurus Transformation

Note that when a thesaurus is used to group medium frequency terms, one obtains broader thesaurus classes. The thesaurus transformation then transforms individual terms with positive discrimination values into broader terms with negative

discrimination values. Even worse is the case when broad terms are included in a thesaurus, because the terms starting out with a negative discrimination value become even worse after the thesaurus transformation. Thus a thesaurus class groupng terms such as

minicomputer
microcomputer
Apple (computer)
Commodore (computer)

operates satisfactorily assuming that all these terms are low-frequency terms. If on the other hand, a class includes both "computer" as well as "minicomputer", the result is necessarily poor, because the high-frequency term "computer" is replaced by a thesaurus class of even greater frequency. Queries about minicomputers are then liable to retrieve documents about computers.

Various semi-automatic or automatic thesaurus construction methods are described in the literature. [5,6]

a) One can take the low-frequency terms exhibiting near-zero discrimination values and group them into similarity classes manually, or rather intellectually.

b) Alternatively the term grouping process can be performed automatically using an automatic clustering process that assembles into common classes terms that occur in similar contexts. The context chosen for term grouping purposes can be global--for example by grouping into common classes terms which co-occur sufficiently often in the documents of a collection; or the context can be local by grouping terms that co-occur in the small subset of documents that are jointly retrieved in response to certain queries.

When the frequency restrictions are observed and terms of low discrimination value are replaced by thesaurus classes with higher discrimination value, the incorporation of a thesaurus transformation will improve retrieval performance. Formal proofs of the usefulness of a thesaurus can in fact be obtained in some circumstances. [7]

The frequency model shows that the generation and assignment of <u>term phrases</u> instead of single terms has an effect which is precisely the inverse of that of a

thesaurus, in the sense that initially broader entities are replaced by more specific ones. When the term phrases include broad terms with negative discrimination values, better terms with positive discrimination values can then be obtained. On the other hand, when the phrases are used to relate terms that on their own are already sufficiently specific, a worse result could be obtained because entities with a positive DV could be transformed into terms of zero DV. The right-to-left phrase transformation is illustrated by the frequency spectrum of Fig. 3. [8,9]

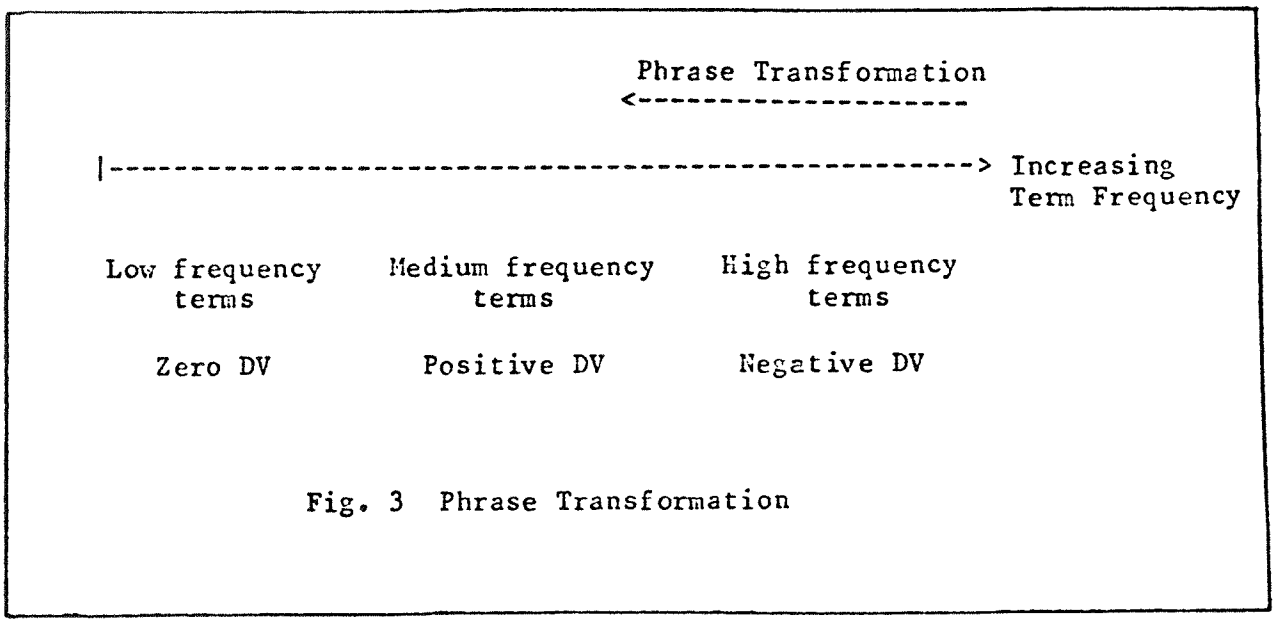

Fig. 3 Phrase Transformation

The phrase formation process should affect mostly the negative discriminators. That is, care must be taken that the phases used for content specification do not appear too far left in the frequency spectrum. For this reason an appropriately loose phrase construction system must be used:

a) At least one component of a phrase should exhibit a negative discrimination value;

b) The terms included in a phrase should co-occur in the same documents, or possibly in the same sentences of documents.

When more stringent phrase formation criteria are used needless distinctions will be made between individual word occurrence patterns. The phrases then become overspecific and recall losses are unavoidable. Consider as an example a document containing the sentence

"an effective retrieval system is useful to people in need of information."

Since "information" is a high-frequency term with a negative discrimination value, a simple phrase formation process based on word co-occurrence properties may correctly produce the phrase "information retrieval." If on the other hand a close syntactic relationship were required between phrase components before a phrase could actually be assigned, the phrase "information retrieval" would be rejected in the earlier example, and the corresponding document might not be retrievable when wanted. [10]

The thesaurus and phrase formation procedures can be added to the simple indexing process outlined earlier in Table 1. An enhanced indexing chart is presented in Table 2.

1.	Identify individual text words	see Table 1
2.	Use stop list to delete common function words	
3.	Use automatic suffix stripping to produce stems	
4.	Compute term discrimination value for all word stems	
5.	Use thesaurus class replacement for all low-frequency terms with near-zero discrimination values	
6.	Use phrase formation process for all high frequency terms with negative discrimination values	
7.	Compute IDF weighting function for all terms	
8.	Assign to each document the corresponding single terms, term phrases and thesaurus classes with IDF weights.	

Enhanced Automatic Indexing Strategy with
Thesaurus and Phrase Assignment
Table 2

The inverse document frequency (IDF) weight of a term in a document was defined earlier as the document frequency of the term divided by the collection frequency. Since a thesaurus class is made up of low frequency terms, the IDF weighting factor for a thesaurus class may be generated as the sum of the document frequencies of the individual terms in the class divided by the sum of the individual collection frequencies. Correspondingly, for phrases constructed from high-frequency components, the IDF weight is computed as the average document frequency of the phrase components divided by the average collection frequency.

5. User System Interaction

At the present time nearly all operational retrieval systems offer on-line user services. The queries are then introduced by using a terminal device, and answers are received while the customer is waiting. In principle, an interactive retrieval system can then be instituted where information obtained from the users during the search process serves for the construction of improved query formulations. In the well-known relevance feedback process, relevance assessments obtained from the users for certain previously retrieved documents are used to construct new query statements which resemble the items previously identified as relevant and differ from the items identified as nonrelevant. [11,12]

When relevance assessments are available from the users for certain documents with respect to certain queries, it also becomes possible to use refined term weighting systems. In computing the previously mentioned IDF weighting functions, no distinctions are made among term occurrences in different kinds of documents. That is, the occurrence of a term in a nonrelevant document is considered as important as an occurrence in a relevant one. When document relevance information is available as it might be in an interactive retrieval environment for certain documents retrieved early in a search, a term relevance factor can be computed for the terms as a function of the proportion of relevant documents in which they occur. In particular, if the term relevance L_j is defined as

$$L_j = \frac{\text{Proportion of relevant items in which } T_j \text{ occurs}}{\text{Proportion of nonrelevant items in which } T_j \text{ occurs}},$$

then a term with a high L_j factor should be able to retrieve additional relevant documents similar to those already seen. In some circumstances, L_j is known to represent an optimal term weighting function for retrieval purposes. [13,14]

The term relevance factor can be used as a weighting function attached to the document terms. The weight of term T_j assigned to document D_i can then be defined as

$$TERMREL_{ij} = (\text{Document Frequency of } T_j \text{ in } D_i) * (L_j).$$

Alternatively, the term relevance factor may be used to construct effective query formulations. [15,16] One possibility in this connection consists in combining the relevance feedback process with the term relevance computation. The following procedure may be used: [17]

1) Start with an initial query

$$Q = (q_0, q_1, \ldots, q_m)$$

where q_i represents query term i.

2) Assign inverse document frequency weights to the query terms, producing

$$Q = (idf_0, idf_1, \ldots, idf_m)$$

where idf_i is defined as the reciprocal of the collection frequency of q_i (that is, 1/collection frequency).

3) Process the query against the document collection and let user identify some retrieved documents as relevant.

4) Choose terms included in the relevant documents for addition to the query.

5) Compute a term relevance factor for all query terms based on the occurrence properties of the terms in the relevant retrieved and the nonrelevant retrieved documents; that is

$$rel_j = \frac{\text{Proportion of relevant retrieved document with term j}}{\text{Proportion of nonrelevant retrieved documents with termj}}$$

6) Construct a new expanded query consisting of the initially available terms plus the added terms $m+1, m+2, \ldots, n$ chosen from the previously retrieved relevant documents; the terms weights are defined as a combinaton of the original idf weights plus the newly computed relevance weights. That is

$$Q_{new} = \alpha(idf_0, idf_1, \ldots, idf_m, 0, 0, \ldots 0) + \beta(rel_0, rel_1, \ldots, rel_m, rel_{m+1}, \ldots, rel_n).$$

7) The parameters α and β are constants where $\alpha + \beta = 1$ and the value of β is used to determine the importance attached to the feedback process. Initially $\alpha = 1$ and $\beta = 0$; as more and more relevant and retrieved items are identified, the value of α is reduced as the value of β grows.

The foregoing process can be repeated several times by operating with Q_{new}, retrieving additional documents, and constructing updated queries for each iteration. As more relevant documents are identified, the estimated term relevance factor rel_j may in time approach the true value L_j. A substantial amount of work has been devoted in recent times to the construction of good methods for estimating the term relevance. [17-21]

6. A Blueprint for Automatic Indexing

The final blueprint combines the term extraction and weighting processes of Table 1, the term grouping procedures based on thesaurus and phrase formation of Table 2, and the use of relevance factors outlined in the previous section. A summary appears in Table 3. In practice, the indexing process can be truncated to include only the term extraction and IDF weighting methods plus the standard relevance feedback procedure. The thesaurus, phrase, and term relevance computations may be added as desired. An indexing strategy based on this outline should produce a high order of effectiveness and outperform alternative manual or semi-automatic indexing methods.

1.	Identify individual text words	Table 1
2.	Use a stop list to delete common words	
3.	Use suffix stripping to produce word stems	
4.	Replace low-frequency terms by thesaurus classes	Table 2
5.	Replace high-frequency terms by phrases	
6.	Compute IDF weights for all single terms, pnrases and thesaurus classes	
7.	Compare query statements with document vectors	
8.	Identify some retrieved documents as relevant and nonrelevant to the query	
9.	Compute term relevance factors based on available relevance assessments	
10.	Construct new queries with added terms from relevant documents and term weights based on combined IDF and term relevance weights	
11.	Return to step 7	

Automatic Indexing with Relevance Feedback

Table 3

References

[1] H.P. Luhn, A Statistical Approach to Mechanized Encoding and Searching of Literary Information, IBM Journal of Research and Development, Vol. 1, No. 2, October 1957, pp. 309-317.

[2] K. Sparck Jones, A Statistical Interpretation of Term Specificity in Retrieval, Journal of Documentation, Vol. 28, No. 1, March 1972, pp. 11-21.

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

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

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

[6] M.E. Lesk, Performance of Automatic Information Systems, Information Storage and Retrieval, Vol. 4, 1968, pp. 201-218.

[7] C.T. Yu, G. Salton and M.K. Siu, Effective Automatic Indexing Using Term Addition and Deletion, Journal of the ACM, Vol. 25, No. 2, April 1978, pp. 210-225.

[8] G. Salton and A. Wong, On the Role of Words and Phrases in Automatic Text Analysis, Computers and the Humanities, Vol. 10, 1976, pp. 69-87.

[9] G. Salton, C.S. Yang and C.T. Yu, Contributions to the Theory of Indexing, Information Processing 74, North Holland Publishing Company, Amsterdam, 1974, pp. 584-590.

[10] G. Salton, Automatic Phrase Matching, in Readings in Automatic Language Processing, D.G. Hays, editor, American Elsevier Publishing Company, New York, 1966.

[11] J.J. Rocchio Jr., Relevance Feedback in Information Retrieval, in The Smart System--Experiments in Automatic Document Processing, G. Salton, editor, Prentice Hall, Englewood Cliffs, New Jersey, 1971, Chapter 14.

[12] G. Salton, Relevance Feedback and the Optimization of Retrieval Effectiveness, in The Smart System-Experiments in Automatic Document Processing, G. Salton, editor, Prentice Hall, Englewood Cliffs, New Jersey, 1971, Chapter 15.

[13] S.E. Robertson and K. Sparck Jones, Relevance Weighting of Search Terms, Journal of the Am. Soc. for Info. Science, Vol. 27, No. 3, May-June 1976, pp. 129-146.

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

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

[16] K. Sparck Jones, Search Term Relevance Weighting--Some Recent Results, Journal of Info. Science, Vol. 1, 1980, pp. 325-332.

[17] H. Wu and G. Salton, Estimation of Term Relevance Weights Using Relevance Feedback, Department of Computer Science, Cornell Univesity, Ithaca, New York, 1981.

[18] 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, January 1982.

[19] C.J. van Rijsbergen, A Theoretical Basis for the Use of Cooccurrence Data in Information Retrieval, Journal of Documentation, 1977, Vol. 33, pp. 106-119.

[20] D.J. Harper and C.J. van Risjbergen, An Evaluation of Feedback in Document Retrieval Using Co-Occurrence Data, Journal of Documentation, 1978, Vol. 34, pp. 189-216.

[21] C.J. van Rijsbergen, D.J. Harper and M.F. Porter, The Selection of Good Search Terms, Information Processing and Management, Vol. 17, 1981, pp. 77-91.