Chapter 3

Text Analysis and Automatic Indexing

0 PREVIEW

In the first two chapters of this book the design and operations of existing information retrieval systems have been presented. Of all the operations required in information retrieval, the most crucial and probably the most difficult one consists in assigning appropriate terms and identifiers capable of representing the content of the collection items. This task, known as indexing, is normally performed manually by trained experts. In modern environments the indexing task can be performed automatically. This chapter is concerned with the techniques used for automatic indexing, and with the effect and performance of these techniques.

The basic indexing task is first described, followed by a comparison of manual and automatic indexing. Basic techniques are then examined for choosing good content terms and for assigning weights to the terms according to their presumed value for content identification. A simple automatic indexing procedure is then suggested, as well as refinements consisting of the use of term phrases and thesaurus classes. The use of linguistic and probabilistic techniques in automatic indexing is also briefly introduced. Finally, evaluation output is included to demonstrate the effectiveness of the proposed indexing techniques applied to small sample collections.

This chapter includes some technical material. The reader is urged to follow the difficulty indications in mastering the theories underlying the automatic indexing process.

1 INDEXING ENVIRONMENT

Of all the procedures normally used in a document processing environment, the most important and also the most difficult to carry out are the *analysis* operations consisting of the assignment to the bibliographic items of terms or identifiers capable of representing document content. In principle, a document analysis process is redundant if the document collection is small enough to permit the scanning of the full text of all items whenever a request for information is received. In practice, such a solution is too time-consuming and too expensive. Hence it is customary to characterize each item by assigning a short description, or profile, to the item which can be used to obtain access whenever the item is wanted. In standard library environments, the analysis operations are known variously as cataloging, classification, indexing, and abstracting.

The document profile fulfills the dual role of representing a given document by providing a short-form description, and also of describing the document content; the profile is therefore often divided into two parts consisting first of objective information relating to the data external to the document text itself, such as author name, publisher, and date and place of publication; and second of identifiers specifically describing the information contained in the document. In conventional libraries the choice of objective document identifiers is known as author/title or descriptive cataloging, whereas the assignment of the content information is termed subject cataloging. As explained in Chapter 1, the content description of an item consists of a call number chosen from a hierarchically organized systematic list, and of additional subject headings that may be represented by relatively free language words and phrases. Each subject heading assigned to a given item may be entered onto a separate catalog card, and the resulting collection constitutes the library catalog. In a modern document processing environment the conventional library catalog does not exist as a physical entity. Rather the profiles are collected into a data base. In this case, the content analysis operations are then collectively known as indexing. When the assignment of the content identifiers is carried out with the aid of modern computing equipment, the operation becomes *automatic indexing*.

It would be nice to think that the choice of the *objective* document identifiers presents no difficulties. In fact, elaborate cataloging standards have had to be defined specifying what author names are acceptable and how the author/title information is to be entered into the catalog. The study of existing author cataloging rules presents many challenges which deserve to be critically examined [1,2]. In the present context, however, the most interesting problems arise in connection with the generation and standardization of the content, as opposed to the objective, identifiers. Therefore, this present chapter is concerned with the analysis of document content and in particular with automatic indexing procedures.

The assignment of content identifiers to the information items is designed to fulfill three related purposes [1]:

1 To allow the location of items dealing with topics of interest to the user
2 To relate items to each other, and thus relate the topic areas, by identifying distinct items dealing with similar, or related, topic areas
3 To predict the relevance of individual information items to specific information requirements through the use of index terms with well-defined scope and meaning

The methods used to accomplish these aims depend on the particular indexing environment in which the operations take place.

The first distinction to be made is that between *manual* and *automatic indexing*. Historically, the analysis operations have been carried out manually—maybe one should say intellectually—by "subject experts." To this day, manual indexing is the rule rather than the exception in most operational environments. A variety of aids are made available to the indexer to control the indexing process including terminology lists, instruction manuals, and specially structured worksheets to record the indexing products. "Scope notes" may also be used to define the meaning and interpretation of each of the allowable index terms. Obviously, lists of definitions and scope notes in natural language text form are not easily incorporated into an automatic indexing system.

The second distinction to be made is that between *controlled* and *uncontrolled indexing terms*. Many experts feel that an uncontrolled indexing vocabulary which in principle can include the whole variety of the natural language introduces too many opportunities for ambiguity and error. Hence, a limited indexing language is often advocated in which the terms available for content identification are rigidly controlled. This permits the control of spelling and the elimination of synonyms by referring to unique accepted terms for each synonym class, and by identifying semantically related terms. The use of controlled terms guarantees retrieval of appropriately marked items when the correct search terms are known, but it also normally implies that trained intermediaries are needed to formulate the query statements.

A third problem relates to the type of vocabulary used for indexing purposes. A distinction is made between the use of *single terms* to characterize document content, as opposed to the use of *terms in context* where relationship indicators may be available to connect several identifiers, and the basic units may consist of compound entries and phrases. In the single term mode, the content identifiers, known as index terms, keywords, or descriptors, are represented by individual words used to express the concepts included in each document. Then each document is characterized by a collection of individual terms. The terms are eventually combined, or "coordinated," to form topic descriptions when the search requests are formulated. This process is known as "postcoordination." On the other hand, when compound terms are utilized for indexing purposes, consisting of phrases possibly including nouns, adjectives, prepositions, and a variety of relationship indicators, the process is called "precoordination." For example, indexing this book under "automatic information retrieval" is a use of precoordination; indexing it separately under each term and finding it in response to a query such as INFORMATION AND RETRIEVAL represents postcoordination.

In many manual indexing situations where trained experts are involved, the use of controlled indexing languages using precoordinated compound terms is preferred. Automatic indexing systems, on the other hand, often use single terms because the automatic assignment of effective single terms is well understood; the terms are then combined by postcoordination at search time only. No matter what indexing environment is preferred, it is always necessary to take into account two characteristics of the indexing products known as "exhaustivity" and "specificity." Exhaustivity refers to the degree to which all the concepts and notions included in a document are recognized in the index descriptions. The more exhaustive the indexing the higher may be the proportion of the relevant items that can be retrieved, because all the various aspects of the subject matter are then properly recognized. Specificity, on the other hand, refers to the generic level of the index terms used to characterize the document content. If the indexing vocabulary is very specific, and if narrowly defined terms are assigned to the bibliographic items, a large proportion of the nonrelevant items may be properly rejected when the documents to be retrieved are determined.

In many retrieval environments it is customary to measure the effectiveness of retrieval by using two parameters for each search known respectively as "recall" and "precision." Recall measures the proportion of relevant information actually retrieved in response to a search (that is, the number of relevant items actually obtained divided by the total number of relevant items contained in the collection), whereas precision measures the proportion of retrieved items actually relevant (that is, the number of relevant items actually obtained divided by the total number of retrieved items). These evaluation measures are examined in more detail in Chapter 5. The use of exhaustive indexing and a specific indexing language is believed to lead to high recall as well as high precision.

The dual characteristics of exhaustivity and specificity are sometimes subsumed under the notions of *deep* and *shallow* indexing. Deep indexing implies both high exhaustivity and specificity and hence a good retrieval performance. Shallow indexing, on the other hand, is produced by using a few broad terms to characterize each document. In these circumstances, the retrieval performance may be expected to suffer somewhat, but the indexing task may be performed more rapidly and more economically.

2 MANUAL AND AUTOMATIC INDEXING

Before turning to a description of automatic indexing methods, it may be useful to summarize briefly some of the conventional manual indexing practices. In most situations a controlled indexing language is used in which a single standard term or phrase represents a wide variety of related terms and descriptions. Thus, if the standard entry "oscillation" is specified in the accepted indexing terminology, alternative related expressions such as "vibration," "undulation," "pulsation," "swing," and "rolling" are replaced by "oscillation" when the documents are indexed and the search requests are formulated. To facilitate the interpretation of the indexing vocabulary and the retrieval of relevant information, the elements of the controlled vocabulary may be used in context, through precoordination of terms. Thus, instead of assigning single terms such as "dyes," "solvents," or "spectra," many systems specify complex indexing entries such as "dyes, spectra, effect of solvents" or "solvents, effect on spectra of dyes" [2].

When precoordinated, controlled indexing languages are manually assigned, it is necessary to abide by the rules relating to the degree of desirable context. That is, the number of related terms that should preferably be used; the order in which the associated terms should be listed; and also the type of relationship indication to be used between the components of an indexing entry are all specified. Thus, the indexing products may be restricted to short phrases or may be of sentence length equivalent to a complete document title. The listing order may prescribe that a thing or object be entered before any action performed on the object, which in turn should be listed before any instrument used in the action. Thus, an entry might have to be specified as "coal, production" and not "production, coal."

Among the relation indicators accepted in various systems are the standard natural language prepositions and conjunctions, as well as *links* to indicate connections between two or more terms in a description, and *roles* to specify the function of particular terms in an indexing description. Thus if one were interested in the hardness of copper and the conductivity of titanium, a link between the first two terms and another one between the last two would relate each substance to the relevant property. Typical "roles" performed by index entry components are "action," "instrument," "object," "subject," etc. Prepositions and conjunctions that are sometimes ambiguous could be used with some types of free-language indexing; links and roles on the other hand are normally defined precisely and used with controlled vocabulary indexing.

To solve the intellectual problems of index language design, and aid the indexer in the term assignment and the searcher in the formulation of search requests, a variety of vocabulary lists and terminology descriptions may be used, including in particular thesauruses that contain lists of equivalent and related terms for each standard thesaurus entry. Hierarchical dictionaries may be available containing general term arrangements capable of identifying broader and narrower terms for the various dictionary entries. The following types of cross references are often included in the existing terminology descriptions:

1 "See" references which identify the standard entry for terms not accepted by the indexing language ("aircraft, see airplanes")

2 "See also" references, sometimes also designated RT (related terms), which provide references between groups of related terms ("accidents, see also collisions, hazards, safety, survival")

3 References to generically broader terms, sometimes designated BT ("conversion coating, BT coating")

4 References to generically narrower terms, sometimes designated NT ("cooling, NT conduction cooling, convection cooling, evaporation cooling")

The terminology lists may take a variety of forms, including in particular alphabetical term arrangements in which the entries and subentries within entries are alphabetically arranged (Table 3-1a). Alternatively systematic, hierarchical term arrangements in tree form can be used where indentation on the page denotes the generic level (in the excerpt of Table 3-1b, "traffic potential" is generically inferior to "traffic," which in turn is inferior to "airlines"). In addition to formal vocabulary arrangements, it is often convenient to maintain lists of index entries derived from the formal thesaurus together with references to the documents indexed by the corresponding entries. To facilitate the consultation of these indexes, it is customary in many cases to repeat each entry

Airlines.	Airplanes
—— Certification	—— Cargo
—— Depreciation	—— Convertible
—— Economics	—— Light
—— Employees	
—— Fares	

(a)

- Airlines
  - Traffic
    - Traffic potential
  - Finance
    - Accounting
  - Operation
  - Equipment

(b)

Peas, deficiency of copper and zinc in
Copper, deficiency in peas
Zinc, deficiency in peas
Deficiencies of copper and zinc in peas

(c)

Table 3-1 Conventional dictionary formats. (a) Excerpt from alphabetical terminology. (b) Excerpt from systematic terminology. (c) Multiple entry articulated term arrangement.

several times by changing the "lead term," that is, the term used to gain entry into the index. Various strategies are used to construct the set of multiple entries, known as rotation, cycling, chaining, permutation, and so on. In each case, the aim is to furnish a separate entry into the index for each standard as well as each related term in a given compound expression. An example is included in Table 3-1c where the various entries (under "peas," "copper," "zinc," and "deficiencies") are all assumed to refer to the same document or documents. A similar strategy is used in the so-called permuted title indexes which are a part of many modern automatic indexing environments.

An evaluation of the performance of standard subject indexes shows that the use of full index term context leads to a much better precision performance than the lack of context exemplified by the use of single term entries. Increased term specificity and a better understanding of the meaning of the index entries may suppress many nonrelevant items that would normally be retrieved if entries devoid of context were used. On the other hand, when the various forms of context are compared with each other, such as the use of rotated entries and of function words to designate term relationships, few differences in retrieval effectiveness are detected [3].

The foregoing considerations make it plain that much can, in principle, be gained by using a sophisticated indexing product in finding useful documents, rejecting extraneous items, and determining the potential importance of the stored information items. These gains are dependent on the quality, accuracy, and consistency of the indexing performance. Not only must indexers be intimately aware of the available indexing vocabularies and practices, but they should also have knowledge of the collection characteristics and of the type of user queries the system may be expected to process in the future. Furthermore, the performance of the various indexers and searchers that must participate in most operational environments ought to be sufficiently consistent to guarantee that similar documents are identified by comparable indexing entries.

In practice one finds that accuracy and consistency are difficult to maintain. The situation demands a good deal of sophistication, training, and experience on the part of the indexing personnel. But more often than not, the resulting index entries are insufficiently exhaustive, omitting relevant entries, or lacking in specificity. The former produces recall losses, while the latter may lead to recall as well as precision deficiencies [4]. A lack of indexing consistency also produces difficulties in detecting similarities among queries and documents [5,6]. Therefore, the potential advantages of strictly controlled, manually applied indexing languages may be largely illusory.

An uncontrolled, natural language indexing system that is applied automatically exhibits substantial advantages. A natural language system, when properly used, can be specific in the sense that the language may provide just the right kind of expression to denote each particular concept. Furthermore, the natural language is the language of discourse. The authors of documents as well as users of information systems are accustomed to the language, and even the human indexers and other subject specialists are likely to feel comfortable dealing with a natural language system. For this reason natural language indexing may be carried out more rapidly, and hence more cheaply than indexing based on a controlled vocabulary and precoordinated terms.

On the other hand, in a natural language indexing system where a controlled thesaurus is not immediately available to make the distinction between acceptable and forbidden index terms, some way must be found to supply the synonymous and related terms that are available in a controlled indexing environment. Various approaches have been used for this purpose. Thus, exhaustive indexing products may be obtained by choosing a wide variety of different index terms for assignment to queries and documents. Term relationships can be added, for example, by constructing term phrases instead of individual terms; and the precoordinated indexing entries available with controlled term systems can be replaced by juxtaposing individual terms when the search requests are formulated. Nevertheless, the generation of effective indexing products remains a major problem in a natural language indexing system.

It has been claimed that automatic, free language indexing products are necessarily inferior to manual systems because automatic systems are *derivative* in the sense that the original document and query texts must serve as principal inputs for the indexing operation. It becomes necessary in these circumstances to wrestle with the peculiarities of the languages used by individual authors, and to worry about unusual terminologies and expressions. Any assertions concerning the inadequacy of automatic indexing can also be bolstered by demonstrations designed to show that the results of particular automatic indexing procedures will fail to pass any rational test carried out by independent human observers [7]. In so doing, one forgets that the results of manual indexing are also influenced by the terminology contained in individual documents, no matter how much control may be provided by the auxiliary indexing aids.

In any case, the justification of any indexing technique ultimately lies in the retrieval results obtained. Substantial evidence now indicates that simple automatic indexing methods are fast and inexpensive, and produce a recall and precision performance at least equivalent to that obtainable in manual, controlled term environments. A wide variety of automatic indexing methods are examined in the remainder of this chapter, and performance evaluation data are given for some of the proposed methods.

3 AUTOMATIC TERM EXTRACTION AND WEIGHTING

A General Considerations

The indexing task consists first of assigning to each stored item terms, or concepts, capable of representing document content, and second of assigning to each term a weight, or value, reflecting its presumed importance for purposes of content identification. The first and most obvious place where appropriate content identifiers might be found is the text of the documents themselves, or the text of document titles and abstracts. This section is thus concerned with methods for the extraction of content terms from documents and document excerpts and with the assignment of term weights in order of term importance.

Most automatic indexing efforts start with the observation that the frequency of occurrence of individual word types (that is, of distinct words) in natural language texts has something to do with the importance of these words for purposes of content representation. Specifically, if all words were to occur randomly across the documents of a collection with equal frequencies, it would be impossible to distinguish between them using quantitative criteria. In fact, it has been observed that the words occur in natural language text unevenly. As a result of this, classes of words are distinguishable by their occurrence frequencies. To quote from H.P. Luhn, one of the pioneers in automatic indexing [8]:

> The justification of measuring word significance by use-frequency is based on the fact that a writer normally repeats certain words as he advances or varies his arguments and as he elaborates on an aspect of a subject. This means of emphasis is taken as an indicator of significance. . . .

In fact, it is known that when the distinct words in a body of text are arranged in decreasing order of their frequency of occurrence (most frequent words first), the occurrence characteristics of the vocabulary can be characterized by the constant rank-frequency law of Zipf:

$$\text{Frequency} \cdot \text{rank} \simeq \text{constant} \qquad (1)$$

That is, the frequency of a given word multiplied by the rank order of that word will be approximately equal to the frequency of another word multiplied by its rank [9]. The law has been explained by citing a general "principle of least effort" which makes it easier for a speaker or writer of a language to repeat certain words instead of coining new and different words. The least-effort principle also accounts for the fact that the most frequent words (those with the lowest ranks) tend to be short function words (and, of, but, the, etc.) which are easy to coin and whose cost of usage is small.

The law has been verified many times using text materials in different areas. A short illustration is contained in Table 3-2 [10]. A typical graph showing the cumulative fraction of word usage in natural language texts is shown in Fig. 3-1. It may be seen from the figure that the most frequent 20 percent of the text words account for some 70 percent of term usage.

Using the *Zipf law* of expression (1) as a starting point, it is now possible to derive word significance factors based on the frequency characteristics of individual words in document texts. An early proposal was based on the following general consideration [8]:

1 Given a collection of n documents, calculate for each document the frequency of each unique term in that document. This is the frequency of term k in document i, or FREQ_{ik}.

Table 3-2 Illustration of Rank-Frequency Law
(Number of Word Occurrences N = 1,000,000)

Rank (R)	Term	Frequency (F)	R · (F/1,000,000)
1	the	69,971	0.070
2	of	36,411	0.073
3	and	28,852	0.086
4	to	26,149	0.104
5	a	23,237	0.116
6	in	21,341	0.128
7	that	10,595	0.074
8	is	10,099	0.081
9	was	9,816	0.088
10	he	9,543	0.095

Adapted from reference 10.

2 Determine the total collection frequency TOTFREQ$_k$ for each word by summing the frequencies of each unique term across all n documents, that is,

$$\text{TOTFREQ}_k = \sum_{i=1}^{n} \text{FREQ}_{ik}.$$

3 Arrange the words in decreasing order according to their collection frequency. Decide on some suitable high threshold value and remove all words with a collection frequency above this threshold. This eliminates high-frequency function words such as those shown in Table 3-2.

4 In the same way, eliminate from consideration low-frequency words. That is, choose some low threshold and remove all words with a collection frequency below this threshold. This deletes terms occurring so infrequently in the collection that their presence does not affect the retrieval performance in a significant way.

5 The remaining medium-frequency words are now used for assignment to the documents as index terms.

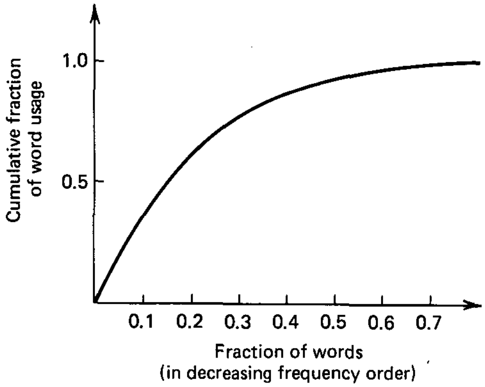

Figure 3-1 Word usage statistics.

Since neither the high- nor the low-frequency terms are good content identifiers, Luhn conjectured that the "resolving power" of the index words extracted from document texts would peak in the middle-frequency range as shown in Fig. 3-2. By "resolving power" is meant the ability of the index terms to identify relevant items and to distinguish them from the nonrelevant material. Among the recommendations made by Luhn for an actual automatic indexing policy, the following may be considered typical [11]:

> A notion occurring at least twice in the same paragraph would be considered a major notion; a notion which occurs also in the immediately preceding or succeeding paragraphs would be considered a major notion even though it appears only once in the paragraph under consideration; notations for major notions would then be listed in some standard order. . . .

There is some evidence that these original ideas are too crude to serve in a practical operational retrieval environment. Thus the elimination of all high-frequency words might produce losses in recall, because the use of broad, high-frequency words for content identification is effective in retrieving large numbers of relevant items. Contrariwise, the elimination of low-frequency terms may produce losses in precision. Another problem is the necessity to choose appropriate thresholds in order to distinguish the useful medium-frequency terms from the remainder. Finally, a question of principle arises concerning the use of *absolute* frequency measures (such as FREQ_{ik} or TOTFREQ_k) for the identification of content indicators. The reason is that a useful index term must fulfill a dual function: on the one hand, it must be related to the information content of the document so as to render the item retrievable when it is wanted (the recall function); on the other hand, a good index term also distinguishes the documents to which it is assigned from the remainder to prevent the indiscriminate retrieval of all items, whether wanted or not (the precision function). Thus a term such as "computer" may never constitute a reasonable term for assignment to a document collection in computing, no matter what its frequency of

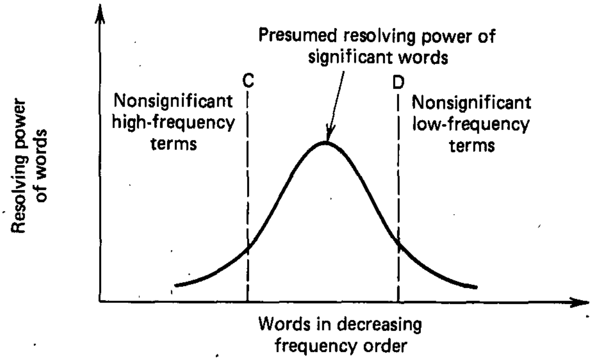

Figure 3-2 Resolving power of significant (medium-frequency) words. (*Adapted from reference 8.*)

occurrence in the documents of the collection, because "computer" is likely to occur in *every* collection item and cannot therefore be used to distinguish the items from each other. This suggests the use of *relative frequency* measures to identify terms occurring with substantial frequencies in some individual documents of a collection, but with a relatively low overall collection frequency. Such terms may then help in retrieving the items to which they are assigned, while also distinguishing them from the remainder of the collection [12,13].

Several term weighting functions have been derived from these basic considerations, including an inverse document frequency function, the signal-noise ratio, and the term discrimination value. These weighting functions are briefly introduced in the remainder of this section.

*B The Inverse Document Frequency Weight

The first possibility consists in assuming that term importance is proportional to the standard occurrence frequency of each term k in each document i (that is, FREQ_{ik}) and inversely proportional to the total number of documents to which each term is assigned. Specifically, one counts the number of documents in which a term k occurs. This produces the document frequency DOCFREQ_k of term k, representing the number of documents to which term k is assigned. A possible measure of the inverse document frequency can now be written as [14]

$$\log_2 \frac{n}{\text{DOCFREQ}_k} + 1 = \log_2(n) - \log_2(\text{DOCFREQ}_k) + 1$$

where n is the number of documents in the collection. For example, in a collection of 1,000 documents, consider the term ALPHA occurring in 100 documents, term BETA occurring in 500 documents, and GAMMA occurring in 900 documents. The inverse document frequency factor will then be 4.322 for ALPHA, 2.000 for BETA, and 1.132 for GAMMA. The emphasis is seen to be placed on the terms exhibiting the lowest document frequencies.

A composite expression measuring the importance, or weight, of term k in a given document i would increase as the frequency of the term in the document, FREQ_{ik}, increases but decrease as the document frequency DOCFREQ_k increases. A possible weighting function is

$$\text{WEIGHT}_{ik} = \text{FREQ}_{ik} \cdot [\log_2 (n) - \log_2 (\text{DOCFREQ}_k) + 1] \qquad (2)$$

This function assigns a high degree of importance to terms occurring in only a few documents of a collection [14].

**C The Signal-Noise Ratio

A related viewpoint suggests using information theory considerations to construct a measure of term importance. In particular, it is known that the information content of a message, or term, can be measured as an inverse function of the probability of occurrence of the words in a given text. Specifically, the higher the probability of occurrence of a word, the less information it contains. The information content of a word is measured as INFORMATION = $-\log_2 p$, where p is the probability of occurrence of the word. For example, if the word ALPHA occurs once in every 10,000 words, its probability of occurrence is 0.0001, and its information is

$$\begin{aligned}\text{INFORMATION} &= -\log_2 (0.0001)\\ &= -(-13.278)\\ &= 13.278\end{aligned}$$

On the other hand, if the word THE occurs once in every 10 words, its probability is 0.1 and its information measure is

$$\begin{aligned}\text{INFORMATION} &= -\log_2 (0.1)\\ &= -(-3.223)\\ &= 3.223\end{aligned}$$

The term information value can be regarded as a measure of reduced uncertainty, in the sense that when terms are assigned as content identifiers to the documents of a collection, knowing a particular term reduces the uncertainty about the document content. Furthermore, the smaller the probability of occurrence of the terms (that is, the greater the term specificity) the larger is the reduction in uncertainty.

By extension, when a document is characterized by t possible identifiers, or terms, each occurring with a specified probability p_k, the average, or expected information (that is, the average reduction in uncertainty about the document) gained by using one of the terms is given by Shannon's formula [15]

$$\text{AVERAGE INFORMATION} = -\sum_{k=1}^{t} p_k \log p_k \tag{3}$$

For example, if the terms ALPHA, BETA, GAMMA, and DELTA are expected to occur with the probabilities 0.5, 0.2, 0.2, and 0.1, respectively, the the average information is

$$\begin{aligned}\text{AVERAGE INFORMATION} &= -[(0.5 \log_2 0.5) + (0.2 \log_2 0.2)\\ &\quad + (0.2 \log_2 0.2) + (0.1 \log_2 0.1)]\\ &= -[(-0.05) + (-0.46) + (-0.46) + (-0.33)]\\ &= 1.3\end{aligned}$$

It is known that the average information is maximized when the occurrence probabilities of the terms are all equal to 1/t for t distinct terms. For example, if ALPHA, BETA, GAMMA, and DELTA are all expected to occur one-fourth of the time, the average information value will be 2 instead of 1.3 as in the earlier example.

By analogy to Shannon's information measure, it is now possible to define the *noise* NOISE_k of an index term k for a collection of n documents

$$\mathrm{NOISE}_k = \sum_{i=1}^{n} \frac{\mathrm{FREQ}_{ik}}{\mathrm{TOTFREQ}_k} \log_2 \frac{\mathrm{TOTFREQ}_k}{\mathrm{FREQ}_{ik}} \tag{4}$$

This measure of noise varies inversely with the "concentration" of a term in the document collection. That is, for perfectly even distributions, when a term occurs an identical number of times in every document of the collection, the noise is maximized. For example, if term k occurs exactly once in each document (all $\mathrm{FREQ}_{ik} = 1$)

$$\begin{aligned}\mathrm{NOISE}_k &= \sum_{i=1}^{n} \frac{1}{n} \log_2 \frac{n}{1} \\ &= \log_2 n\end{aligned}$$

On the other hand, for perfectly concentrated distributions, when a term appears in only one document with frequency $\mathrm{TOTFREQ}_k$, the noise is zero because in that case

$$\begin{aligned}\mathrm{NOISE}_k &= \frac{\mathrm{TOTFREQ}_k}{\mathrm{TOTFREQ}_k} \log_2 \frac{\mathrm{TOTFREQ}_k}{\mathrm{TOTFREQ}_k} \\ &= 1 \log_2 1 \\ &= 0\end{aligned}$$

A relation clearly exists between noise and term specificity, because broad, nonspecific terms tend to have more even distributions across the documents of a collection, and hence high noise. An *inverse* function of the noise might then be used as a possible function of term value [16,17]. One such function, known as the *signal* of term k, is defined as follows

$$\mathrm{SIGNAL}_k = \log_2 (\mathrm{TOTFREQ}_k) - \mathrm{NOISE}_k \tag{5}$$

For the maximum noise case previously discussed (where each FREQ_{ik}) is equal to 1 the SIGNAL is equal to 0, since $\mathrm{TOTFREQ}_k$ in that case equals n. On the other hand, when a term occurs in only one document, a maximum signal of $\log_2 \mathrm{TOTFREQ}_k$ is obtained.

In principle, it is possible to rank the index words extracted from the documents of a collection in decreasing order of the signal value. Such an ordering favors terms that distinguish one or two specific documents (the ones in which the high-signal term exclusively occurs) from the remainder of the collection. Alternatively, the importance, or weight, of term k in document i can be computed as a composite function taking into account FREQ_{ik} as well as SIG-

NAL_k. A possible measure of this type analogous to the term weighting function of expression (2) is

$$\mathrm{WEIGHT}_{ik} = \mathrm{FREQ}_{ik} \cdot \mathrm{SIGNAL}_k \tag{6}$$

It will be seen later that the signal value does not give optimal performance in a retrieval environment.

*D The Term Discrimination Value

Luhn's early proposals were designed to measure the "resolving power" of a term with respect to a document by using the frequency of occurrence of the term in the document. Another approach is to compute the *discrimination value* of a term. This measures the degree to which the use of the term will help to distinguish the documents from each other [18,19]. Consider, in particular, a collection of documents, and let D_i and D_j represent two documents each identified by a set of index terms. A similarity measure $\mathrm{SIMILAR}(D_i, D_j)$ can be used to represent the similarity between the documents. Typical similarity measures generate values of 0 for documents exhibiting no agreement among the assigned index terms, and 1 when perfect agreement is detected. Intermediate values are obtained for cases of partial agreement.

If the similarity measure is computed for all pairs of documents (D_i, D_j) except when $i = j$, an average value AVERAGE-SIMILARITY is obtainable. This represents the average document-pair similarity for the collection. Specifically,

$$\text{AVERAGE-SIMILARITY} = \text{CONSTANT} \sum_{i=1}^{n} \sum_{\substack{j=1\\ i\neq j}}^{n} \mathrm{SIMILAR}(D_i, D_j) \tag{7a}$$

for some constant [for example, 1/n(n − 1)]. The foregoing expression reflects the *density* of the document space, that is, the degree to which the documents are bunched up in the "space" of documents. When all n documents are identical, $\mathrm{SIMILAR}(D_i,D_j) = 1$ for all document pairs, and AVERAGE-SIMILARITY reaches a maximum:

$$\begin{aligned}\text{AVERAGE-SIMILARITY} &= \text{CONSTANT} \sum_{i=1}^{n} \sum_{\substack{j=1\\ i\neq j}}^{n} 1 \\ &= \text{CONSTANT} \cdot n(n-1)\end{aligned}$$

The space density can be computed more efficiently by constructing an artificial, "average" document $\overline{D}$ as the *centroid*, in which the terms are assumed to exhibit average frequency characteristics, that is, the average frequency of term k is defined as

$$(\text{AVERAGE FREQ})_k = \frac{1}{n} \sum_{i=1}^{n} \mathrm{FREQ}_{ik}$$

The density is then computed as the sum of the similarities of each document with the centroid

$$\text{AVGSIM} = \text{CONSTANT} \sum_{i=1}^{n} \text{SIMILAR}(\overline{\text{D}},\text{D}_i) \tag{7b}$$

Consider now the original document collection with term k removed from all the documents and let $(\text{AVGSIM})_k$ represent the space density in that case. If term k had been a broad, high-frequency term with a fairly even frequency distribution, it is likely that it would have appeared in most document descriptions; therefore, its removal will reduce the average document-pair similarity. This case is clearly unfavorable, because when such a high-frequency term is assigned to the documents, the average similarity will increase and the document space is compressed. On the other hand, if term k had been assigned a high weight in some documents, but not in others, its removal would be likely to increase the average similarity between documents.

The discrimination value DISCVALUE_k can now be computed for each term k as

$$\text{DISCVALUE}_k = (\text{AVGSIM})_k - \text{AVGSIM} \tag{8}$$

Following the computation of DISCVALUE_k for all terms k, the terms can be ranked in decreasing order of the discrimination value DISCVALUE_k. A typical ranking of this type for document collections in three different subject areas appears in Table 3-3. In each case the 10 best discriminators are shown—those whose removal will compress the document space the most—as well as the 10 worst discriminators (other than the previously eliminated common function words whose discrimination values would no doubt be even poorer). It may be seen that the terms at the top of the table are highly specific, whereas the terms at the bottom are much more general. A term such as "flow" occurs in most documents in the Cranfield collection on aerodynamics, accounting for its poor performance as a document discriminator.

For experimental purposes the index terms may be placed into three rough categories according to their discrimination values:

1 The good discriminators with a positive DISCVALUE_k whose introduction for indexing purposes decreases the space density

2 The indifferent discriminators with a DISCVALUE_k close to zero whose removal or addition leaves the similarity among documents unchanged

3 The poor discriminators whose utilization renders the documents more similar, producing a negative DISCVALUE_k

Frequency distributions for three typical terms, one from each category are shown in Table 3-4. It may be seen that the negative discriminator in the rightmost column of the table exhibits a total collection frequency TOTFREQ_k of 527 and a document frequency DOCFREQ_k of 337 for a collection of 450 docu-

Table 3-3 Best and Worst Discriminators for Three Collections
(Cranfield: 424 Documents in Aerodynamics; MED: 450 Documents in Medicine; *Time:* 425 Documents in World Affairs)

Cranfield 424	**MED 450**	***Time* 425**
	a Best discriminators	
1. Panel	1. Marrow	1. Buddhist
2. Flutter	2. Amyloidosis	2. Diem
3. Jet	3. Lymphostasis	3. Lao
4. Cone	4. Hepatitis	4. Arab
5. Separate	5. Hela	5. Viet
6. Shell	6. Antigen	6. Kurd
7. Yaw	7. Chromosome	7. Wilson
8. Nozzle	8. Irradiate	8. Baath
9. Transit	9. Tumor	9. Park
10. Degree	10. Virus	10. Nenni
	b Worst discriminators	
2642. Equate	4717. Clinic	7560. Work
2643. Theo	4718. Children	7561. Lead
2644. Bound	4719. Act	7562. Red
2645. Effect	4720. High	7563. Minister
2646. Solution	4721. Develop	7564. Nation
2647. Method	4722. Treat	7565. Party
2468. Press	4723. Increase	7566. Commune
2649. Result	4724. Result	7567. U.S.
2650. Number	4725. Cell	7568. Govern
2651. Flow	4726. Patient	7569. New

Adapted from reference 19.

ments; the term occurs once in 221 documents, twice in 75 additional documents, three times in 19 documents, four times in 15 documents, and five and six times in 3 and 4 items, respectively. It is not surprising that such a ubiquitous term operates poorly as a discriminator.

The second and third columns of Table 3-4 contain examples of an indifferent discriminator and a good discriminator, respectively. It may be seen that the indifferent discriminator term has a very low document frequency of 16 items out of 450. Its assignment leaves the document space more or less unchanged. The good discriminator in the third column has a document frequency of 61 out of 450 and a total collection frequency of 188.

The document frequency DOCFREQ_k, total collection frequency TOTFREQ_k, and average frequency of each term are shown in Table 3-5 for the 10 best discriminators and 10 worst discriminators in a collection of 852 documents in ophthalmology. In each case the rank of each term in decreasing order of the discimination value is also shown in the table.

The data of Tables 3-4 and 3-5 and related comparisons of the frequency

Table 3-4 Distribution Characteristics of a Typical Term in Each of Three Discrimination Categories
(Collection Size 450)

	Number of documents with corresponding frequency		
Number of occurrences of term k in documents	**Low-frequency term zero DISCVALUE$_k$**	**Medium-frequency term positive DISCVALUE$_k$**	**High-frequency term negative DISCVALUE$_k$**
1	10	26	221
2	3	13	75
3	3	8	19
4	—	4	15
5	—	2	3
6	—	2	4
7	—	—	—
8	—	2	—
9	—	—	—
10	—	—	—
11–15	—	2	—
16–20	—	2	—
21–25	—	—	—
26–30	—	—	—
30+	—	—	—
Total term frequency TOTFREQ$_k$	25	188	527
Total document frequency DOCFREQ$_k$	16	61	337

characteristics of terms with their discrimination value confirm Luhn's original notions that medium-frequency words are to be preferred for assignment as index terms. The discrimination value computation [expression (8)] provides an objective method for determining the frequency thresholds: high-frequency terms with a negative DISCVALUE$_k$ are poor and should not be used directly for indexing purposes; low-frequency terms with a zero DISCVALUE$_k$ may or may not be used—their assignment will not hurt the performance of the retrieval system but may be questioned on efficiency grounds because the storage and manipulation of large numbers of low-frequency terms tends to be expensive; the good discriminators—those with resolving power to use Luhn's terminology—have a positive DISCVALUE$_k$, and they happen to be medium-frequency terms in the collection in which they occur [21,22].

A display of the frequency distribution of good, indifferent, and poor terms in signal-value order [expression (5)] shows that the terms with the best SIGNAL$_k$ values have very low document and collection frequencies. Typical DOCFREQ$_k$ and TOTFREQ$_k$ values for those terms with good signal values are 3, 20; 6, 33; and 2, 9, respectively, showing that the terms with the best "information value" are not those best able to distinguish a substantial number of documents in a collection [18].

Table 3-5 Highest-Ranking Discriminator and Nondiscriminator Terms
(852 Abstracts in Ophthalmology)

Nondiscriminators				Discriminators			
Term	Document frequency	Total frequency	Average frequency	Term	Document frequency	Total frequency	Average frequency
8672. Patient	201	408	2.03	1. Rubella	10	47	4.70
8671. At	194	292	1.51	2. Capillary	19	54	2.84
8670. Use	179	247	1.38	3. Laser	11	32	2.91
8669. Have	194	257	1.32	4. Collagen	12	40	3.33
8668. Retinal	134	275	2.05	5. Cyst	17	42	2.47
8667. Present	184	219	1.19	6. Cholinesterase	6	26	4.33
8666. Has	171	231	1.35	7. Fiber	16	50	3.13
8665. Effect	150	259	1.73	8. Cyclodialysis	4	12	3.00
8664. Result	179	234	1.31	9. Implant	18	36	2.00
8663. Found	174	228	1.31	10. Uveitis	21	45	2.14
8662. Report	141	172	1.22	11. Vessel	36	82	2.28
8661. Occular	125	194	1.55	12. Spray	2	25	12.50

The term discrimination value of expression (8) can be used to compute an importance factor, or weight, for each term in each document of a collection by combining the term frequency factor with the discrimination value. This produces a weighting expression for term k in document i analogous to expressions (2) and (6) for the inverse document frequencies and the signal values as follows:

$$\mathrm{WEIGHT}_{ik} = \mathrm{FREQ}_{ik} \cdot \mathrm{DISCVALUE}_k \qquad (9)$$

The insights concerning the term occurrence frequencies are incorporated into a simple automatic indexing process described in the following section.

4 A SIMPLE AUTOMATIC INDEXING PROCESS

It may be of interest briefly to describe a simple process for the automatic assignment of index terms to the documents of a collection. Such a process must start with the identification of all the individual words that constitute the documents. A problem arising in this connection is the definition of the document to be used. There are many so-called full-text retrieval systems where the full text of the documents is used for indexing purposes. This is true in specialized areas of discourse, for example, law or medicine, where the vocabulary may be specialized and the presence of a particular term, say "tort" in a legal document, has specific connotations [23]. However, the computer storage of the full text of documents is expensive and is rarely possible except as a by-product of automatic typesetting operations. For many practical purposes, it is sufficient to use document excerpts for analysis, such as the titles and abstracts. The available experimental evidence indicates that the use of abstracts in addition to titles brings substantial advantages in retrieval effectiveness. However, the additional utilization of the full texts of the documents appears to produce very little improvement over titles and abstracts alone in most subject areas [24].

Following the identification of the words occurring in the document texts, or abstracts, the high-frequency function words need to be eliminated. These comprise 40 to 50 percent of the text words, and as suggested earlier, these words are poor discriminators and cannot possibly be used by themselves to identify document content. In English, about 250 common words are involved, and it is easy to include them in a dictionary, sometimes called a negative dictionary, or *stop list*. An excerpt from a typical stop list is shown in Table 3-6.

The next step, following removal of stop words, is the identification of "good" index terms and their assignment to the documents of a collection. It is useful first to remove word suffixes (and possibly also prefixes), thereby reducing the original words to word stem form. This reduces a variety of different forms such as analysis, analyzing, analyzer, analyzed, and analysing to a common word stem "analy." The word stem "analy" will have a higher frequency of occurrence in the document texts than any of the variant forms. The generation of word stems, and subsequent identification of common stems, is rela-

Table 3-6 Excerpt from Typical Stop List

A
ABOUT
ACROSS
AFTER
AFTERWARDS
AGAIN
AGAINST
ALL
ALMOST
ALONE
ALONG
ALREADY
ALSO
ALTHOUGH
ALWAYS
AMONG
AMONGST
AN
AND
ANOTHER
ANY
ANYHOW
ANYONE
ANYTHING
ANYWHERE
ARE
AROUND
AS
AT
BE
BECAME
BECAUSE
BECOME
BECOMES
BECOMING
BEEN
BEFORE
BEFOREHAND
BEHIND
BEING
BELOW
BESIDE
BESIDES
BETWEEN
BEYOND
BOTH
BUT
BY
CAN

Excerpted from reference 25.

tively easy to do for many languages (including English) and serves as a recall-enhancing device. When the stems are used as index terms, a greater number of potentially relevant items can be identified than when one of the original full text words is in use.

Several well-known algorithms exist for the removal of word endings, generally based on the use of a list of suffixes followed by the removal of the longest suffix matching any entry on the suffix list [26,27]. An excerpt from a typical suffix list is shown in Table 3-7. In using a suffix removal algorithm it is important to handle various classes of exceptional cases. The following list identifies problems in English:

1 It is desirable to remove the suffix "ability" from computability, or the suffix "ing" from singing; however, the same suffixes should not be removed from the words ability and sing, respectively; problems such as these are normally solved by specifying a minimum stem length that must remain following the suffix removal.

2 Several suffixes may be attached to a single stem; thus effectiveness may be shortened to effective by the removal of "ness," which can in turn be shortened to effect by removal of "ive"; multiple suffixes can be handled either by applying the suffix removal process recursively several times or else by including in the suffix dictionary all multiple suffix entries, and removing longer suffixes in preference to the shorter ones.

3 Various examples of morphological transformations exist in English which may alter the stem of many suffixed words; for example, the word absorb is transformed into absorption when the suffix "tion" is added; similarly hop becomes hopping, relief becomes relieving, and so on. Transformational rules can be set up in order to recode various automatically generated stems fol-

Table 3-7 Exerpt from Typical Suffix List

ABILITIES
ABILITY
ABLE
ABLED
ABLEDLY
ABLENESS
ABLER
ABLES
ABLING
ABLINGFUL
ABLINGLY
ABLY
ACEOUS
ACEOUSLY
ACEOUSNESS
ACEOUSNESSES
ACIDOUS
ACIDOUSLY
ACIES
ACIOUSNESS
ACIOUSNESSES
ACITIES
ACITY
ACY
AE
AGE
AGED
AGER
AGES
AGING
AGINGFUL
AGINGLY
AIC
AICAL
AICALLY
AICALS
AICISM
AICISMS
AICS
AL
ALISATION
ALISATIONAL
ALISATIONALLY
ALISE
ALISED
ALISEDLY
ALISER

Excerpted from reference 25.

lowing suffix removal. A typical rule might state "remove one of double occurrences of b, d, g, l, m, n, p, r, s, t from the end of a generated stem."

4 Finally, a number of additional exceptions which depend on the particular word context must be taken care of, using various context-sensitive rules. For example, a rule for the suffix "allic" specifies a minimum stem length of three and prevents suffix removal after "met" or "ryst"; another rule applicable to the suffix "yl" permits removal only after "n" or "r" [27].

In summary, it is not difficult to implement a suffix removal algorithm producing usable word stems for the vast majority of existing English word forms. A stored suffix list must be used together with a few contextual rules applicable to certain suffixes. A list of transformations to recode some of the generated stems is also necessary, or a stored full word dictionary could be used for that purpose.

After the word stems are generated, it becomes necessary to recognize equivalent stems occurring in the texts and to choose those stems to be used as index terms. The frequency-based techniques can be used to determine the potential usefulness of the remaining word stems. A high standard of performance at modest cost is obtainable by using the inverse document frequency function 1/DOCFREQ$_k$ to obtain a term importance factor. Another possibility consists in using the discrimination DISCVALUE$_k$ or the SIGNAL$_k$. The latter, however, emphasizes term concentration in only a few documents of a collection and should be used only in order to emphasize precision at the expense of recall.

The terms (word stems) with sufficiently high term value factors can be assigned to the documents of the collection either with or without a term weight. When the indexing mode is *binary*, a term that occurs in a document is assigned an implicit weight of 1, no matter what its actual frequency of occur-

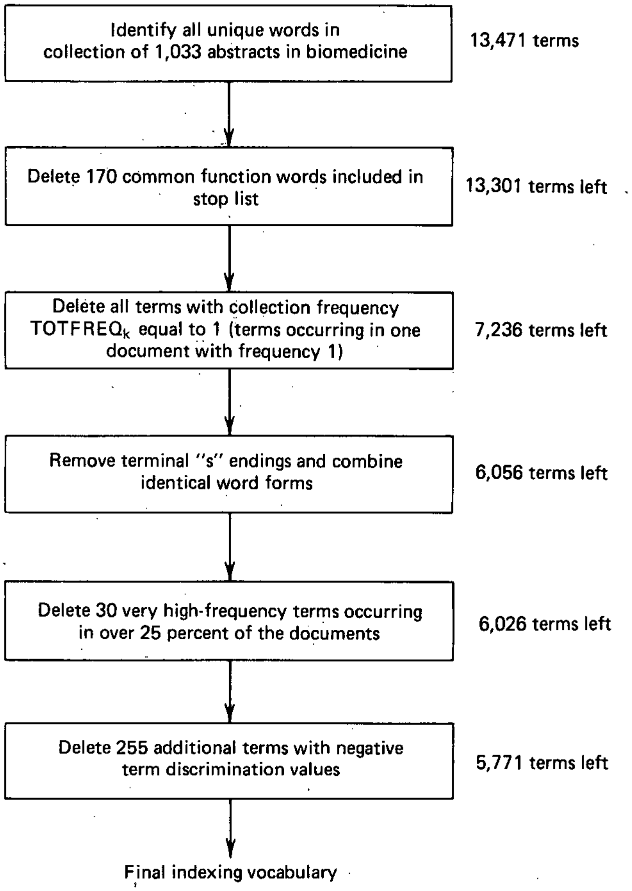
Identify all unique words in collection of 1,033 abstracts in biomedicine
13,471 terms
Delete 170 common function words included in stop list
13,301 terms left
Delete all terms with collection frequency TOTFREQ$_k$ equal to 1 (terms occurring in one document with frequency 1)
7,236 terms left
Remove terminal "s" endings and combine identical word forms
6,056 terms left
Delete 30 very high-frequency terms occurring in over 25 percent of the documents
6,026 terms left
Delete 255 additional terms with negative term discrimination values
5,771 terms left
Final indexing vocabulary

Figure 3-3 Typical term deletion algorithm (data for 1,033 documents in medicine).

rence. In a *weighted indexing* system, a term weight may be used to reflect term importance by using the weighting functions [expressions (2), (6), or (9)] previously described. This produces for each document D_i a *document vector*

$$D_i = \langle d_{i1}, d_{i2}, \ldots, d_{it} \rangle \tag{10}$$

where each d_{ij} is the weight assigned to the jth identifier for document D_i. For example, if there are three terms ALPHA, BETA, and GAMMA, respectively, then

$$D_1 = \langle 2, 4, 0 \rangle$$

means that document number 1 is identified by the term ALPHA with a weight of 2, BETA with a weight of 4, and GAMMA with a weight of 0. The vector length t corresponds to the number of distinct terms assigned to the whole collection, and weights of 0 are assumed for terms not assigned to a given document vector.

It remains to determine what to do with terms whose importance factors are not high enough to make it reasonable to assign them to the documents. In principle such terms can simply be deleted from the identifying vocabulary. A prototype indexing system based on various term deletion methods is reproduced in simplified form in the flowchart of Fig. 3-3. The index term data of Fig. 3-3 are based on the processing of 1,033 document abstracts in medicine. A simplified stemming method was used in which the only recognized suffix is a terminal "s." The indexing vocabulary eventually is reduced to 5,771 stems from the original 13,471 words.

Term deletion methods must be used with caution because the removal of some broad high-frequency terms may produce unwanted recall losses, whereas deletion of certain low-frequency terms reduces indexing exhaustivity and may result in reduced retrieval recall and precision. Instead of deleting the poor discriminators, it may be preferable to improve such terms by turning them into terms with better discrimination properties. This can be done in various ways by using context and term associations, as explained in the next section.

5 AUTOMATIC TERM ASSOCIATION AND USE OF CONTEXT

A Thesaurus Rules

It was seen earlier that some words or word stems extracted from document texts may not function effectively as index terms. This is the case notable for very high-frequency terms that occur in a large proportion of the documents of a collection, and for very low-frequency words which occur very rarely. The question is whether such terms can be transformed into different types of entities that prove more discriminating and better able to reflect document content. The natural language provides a variety of devices for changing the specificity and scope of individual terms: for example, the phrase "term specificity" has a narrower, more specific interpretation than either "term" or "specificity" alone; similarly, the term "computer" has a broader meaning than "minicomputer."

In the preceding discussion, several tools have been described that may be useful for controlling or changing the scope of individual words or terms. Thus, a variety of dictionaries may be available in conventional indexing situations which allow the manual indexer to choose broader or narrower or related terms in addition to or instead of an initially available dictionary entry. A term broadening step was also included in the basic term extraction methods examined in the preceding section in the form of a word stemming process. The stemming process replaces a full text word by a word stem with a broader interpretation.

The basic idea in improving the usefulness of index terms with questionable discrimination properties then consists in using *associations* between terms in the hope of refining or broadening the interpretation of these terms. Many kinds of term associations can profitably be incorporated into an automatic indexing system. The first and most obvious one consists in imitating the manual indexing process by using a *term thesaurus*.

A thesaurus provides a grouping, or classification, of the terms used in a given topic area into categories known as thesaurus classes. As in the manual indexing case, thesauruses can be used for language normalization purposes in order to replace an uncontrolled vocabulary by the controlled thesaurus category identifiers. A thesaurus may broaden the vocabulary terms by addition of thesaurus class identifiers to the normal term lists, thereby enhancing the recall performance in retrieval. Alternatively the thesaurus class identifiers can replace the original term entries in the hope of improving recall and providing vocabulary normalization. When hierarchical relationships are supplied for the entries in a thesaurus in the form of "broader" or "narrower" terms, the indexing vocabulary can be "expanded" in various directions by adding these broader or narrower terms, or certain related terms, as the case may be.

An excerpt of a thesaurus used in an automatic indexing environment for documents in engineering is shown in Table 3-8. The thesaurus class identifiers are represented by identifying "concept numbers" designating the various term classes. Thus, when a document contains the term "superconductivity" or (stem "superconduct"), that term may be replaced by class identifier 415. The same operation could be used for another document, or for a user query, containing the term "cryogenic." Should the document contain "superconductivity" while the query term is "cryogenic," a term match would result through the thesaurus transformation, but not using the original word stems.

Thesauruses may be constructed manually, semiautomatically, and fully automatically. No matter what process is used, two separate problems arise at once:

1 A decision must be made about what terms should be included in the thesaurus.

2 The terms specified for inclusion must be suitably grouped.

To decide what to include, the various term value models described earlier can be used. The discrimination value model specifies, for example, that the most important terms are those with medium document frequency, followed by those with low document frequencies and near zero discrimination values. Since the main purpose of a term classification is the improvement of the recall performance, one concludes that a thesaurus should certainly include a grouping of the low-frequency terms into classes of higher frequency. In addition, a grouping of the medium-frequency good discriminators might also be useful for some purposes, particularly when a high recall performance is wanted. On the other hand, the high-frequency low discriminators might be eliminated altogether.

The following thesaurus construction principles derived in part from the

Table 3-8 Typical Thesaurus Excerpt

408	DISLOCATION JUNCTION MINORITY-CARRIER N-P-N P-N-P POINT-CONTACT RECOMBINE TRANSITION UNIJUNCTION
409	BLAST-COOLED HEAT-FLOW HEAT-TRANSFER
410	ANNEAL STRAIN
411	COERCIVE DEMAGNETIZE FLUX-LEAKAGE HYSTERESIS INDUCT INSENSITIVE MAGNETORESISTANCE SQUARE-LOOP THRESHOLD
412	LONGITUDINAL TRANSVERSE
413	CAPACITANCE IMPEDANCE-MATCHING IMPEDANCE INDUCTANCE MUTUAL-IMPEDANCE MUTUAL-INDUCTANCE MUTUAL NEGATIVE-RESISTANCE POSITIVE-GAP REACTANCE RESIST SELF-IMPEDANCE SELF-INDUCTANCE SELF
414	ANTENNA KLYSTRON PULSES-PER-BEAM RECEIVER SIGNAL-TO-RECEIVER TRANSMITTER WAVEGUIDE
415	CRYOGENIC CRYOTRON PERSISTENT-CURRENT SUPERCONDUCT SUPER-CONDUCT
416	RELAY

earlier indexing models and in part from previously obtained experimental evidence can be enunciated [28]:

1 The thesaurus should include only those terms likely to be of interest for content identification in a subject area (for example, a term such as "hand" might be used in a thesaurus dealing with biology, but it should not be included if its frequency of occurrence is due largely to expressions such as "on the other hand").

2 Ambiguous terms should be coded only for those senses likely to be important in the document collection (at least two thesaurus categories should thus be used for a term such as "field," corresponding on the one hand to the notion of subject area and on the other hand to its technical sense in algebra; no provision need be made to cover the notions of "a patch of land" if the thesaurus deals with the mathematical sciences or related technical fields).

3 In order to obtain good matching characteristics between query and document terms, each thesaurus class should include terms of roughly equal frequency; furthermore, the total frequency of occurrence should be as close to equal for each class as possible, thus ensuring that the probability of producing a match between queries and documents is approximately equal for all thesaurus classes. (If these frequency characteristics are grossly violated—for example, if a high-frequency term such as "computer" is entered into the same class as a more specific term such as "minicomputer"—queries about specific topics will produce general responses, thereby depressing the precision of the search.)

4 Whenever possible, terms with negative discrimination values should be eliminated; even if the size restrictions that control the thesaurus construction do not immediately lead to the elimination of all high-frequency nondiscriminators, the latter are best relegated to thesaurus classes of their own (their classification together with lower-frequency terms would produce low-precision output).

Concerning now the actual thesaurus construction method, a manual thesaurus generation process is an art rather than a science. In recent years, a number of automatic aids have considerably simplified the thesaurus construction task. Thus, given a collection of documents, it is now easy to automatically produce concordances exhibiting the occurrences of all terms in the context in which they occur, arranged in alphabetical order for convenient access. Thus all occurrences of the term "information" would be collected under the letter I, together with contextual information for each occurrence of the term. This makes it possible to determine the placement of each term within a thesaurus class arrangement by collecting in a common class various terms occurring in a given document set in the same context.

An automatically constructed alphabetical arrangement of terms derived from a given document set can in fact function as a kind of thesaurus, and has been widely used in practice to obtain access to document collections. Normally, the terms included in such a listing are the words occurring in the titles of documents. The resulting products are known as keyword-in-context (KWIC) indexes. Alternatively, related term arrangements known as KWAC and KWOC (keyword and context, keyword out of context) are also obtainable. An example of KWIC and KWAC arrangements is shown in Table 3-9. The entries shown in the table are produced by a document (number 3,313) entitled "User Preference in Published Indexes." This title generates four entries: one under I for the term "indexes," two under P for "preference" and "published," and finally one under U for "user."

When aids such as KWIC indexes are used judiciously, and the previously mentioned thesaurus construction principles are applied, the task of building the term classification is simplified. The main intellectual decisions for the actual term grouping process are, however, reached manually.

*B Automatic Thesaurus Construction

A variety of fully automatic thesaurus construction methods are available, based on the use of a set of document vectors of the type shown in expression (10). A document collection is then representable by a matrix such as that in Table 3-10. It was seen earlier that a similarity function SIMILAR(D_i,D_j) re-

Table 3-9 KWIC and KWAC Entries Produced by Document on "User Preference in Published Indexes"

KWIC		
RENCE IN PUBLISHED	INDEXES/ USER PREFE	3,313
HED INDEXES/ USER	PREFERENCE IN PUBLIS	3,313
USER PREFERENCE IN	PUBLISHED INDEXES/	3,313
UBLISHED INDEXES/	USER PREFERENCE IN P	3,313
KWAC		
INDEXES		
USER PREFERENCE IN PUBLISHED INDEXES		3,313
PREFERENCE		
USER PREFERENCE IN PUBLISHED INDEXES		3,313
PUBLISHED		
USER PREFERENCE IN PUBLISHED INDEXES		3,313
USER		
USER PREFERENCE IN PUBLISHED INDEXES		3,313

Adapted from reference 3.

flecting index term similarities can be computed for each document pair (D_i, D_j) by comparing pairs of rows of the document matrix. While the rows of the matrix represent the individual document vectors, the columns identify the term assignments to the documents. That is, a column, j, of the document vector matrix reflects the assignment of $TERM_j$ to the documents of the collection. The vector comparison process previously used to compute the density of the document space [see expression (7)] can also be used to obtain a similarity measure between pairs of columns $SIMILAR(TERM_k, TERM_h)$, reflecting the similarities between $TERM_k$ and $TERM_h$. Given term vectors of the form $TERM_k = (t_{1k}, t_{2k}, \ldots, t_{nk})$, where t_{ik} indicates the weight or value of $TERM_k$ in document i and assuming n documents in the collection, a typical similarity measure may then be defined as

$$SIMILAR(TERM_k, TERM_h) = \sum_{i=1}^{n} t_{ik}\, t_{ih} \qquad (11)$$

Table 3-10 Matrix of Document Vectors

$$\begin{array}{c|cccc} & T_1 & T_2 & \cdots & T_t \\ \hline D_1 & d_{11} & d_{12} & \cdots & d_{1t} \\ D_2 & d_{21} & d_{22} & \cdots & d_{2t} \\ \vdots & \vdots & & & \vdots \\ D_n & d_{n1} & d_{n2} & \cdots & d_{nt} \end{array}$$

or, using a normalization factor to limit the computed results to values between 0 and 1,

$$\text{SIMILAR}(\text{TERM}_k,\text{TERM}_h) = \frac{\sum_{i=1}^{n} t_{ik}\, t_{ih}}{\sum_{i=1}^{n} (t_{ik})^2 + \sum_{i=1}^{n} (t_{ih})^2 - \sum_{i=1}^{n} t_{ik}\, t_{ih}} \tag{12}$$

When all pairs of distinct columns of the matrix of Table 3-10 are compared with each other, a *term-term association* matrix T is constructed in which the element located in row k and column h equals SIMILAR(TERM_k,TERM_h). A sample term-term association matrix is shown in Table 3-11.

A variety of *automatic classification* or clustering methods can now be used to construct classes of similar terms (equivalent to thesaurus classes) by collecting in a common class all terms whose similarity coefficients SIMILAR are sufficiently large [29,30]. Automatic clustering methods are covered in detail in Chapter 6. Many different methodologies are available. For example, the *single-link* process collects in a single class all items TERM_k such that the similarity between TERM_k and at least one other member of the same class exceeds some threshold. In the *clique* process the similarity between TERM_k and *all* other members of the same class must exceed the stipulated threshold.

In the single-link or clique methods, the term classes are constructed from the beginning starting from the term assignments to the documents of a collection. A number of classification methods assume the prior existence of term classes, and proceed by refining the initial state of the classification. Various possibilities exist for defining such an initial term grouping:

1 A given term class may be defined as the set of terms assigned to a particular document, or document set; this generates a number of initial term classes equal to the number of documents used as starting sets.

2 A term class might also be defined as the terms contained in the set of relevant documents retrieved in response to certain user queries; here the number of initial term classes is equal to the number of starting user queries for which relevance information is available.

For each existing class, a centroid TERM-CENTROID = $\langle \bar{t}_1, \bar{t}_2, \ldots, \bar{t}_m \rangle$ can then be defined as the average vector for the term vectors of that class.

Table 3-11 Term-Term Similarity Matrix

$$\begin{array}{c|cccc} & T_1 & T_2 & \cdots & T_t \\ \hline T_1 & s(T_1,T_1) & s(T_1,T_2) & \cdots & s(T_1,T_t) \\ T_2 & s(T_2,T_1) & s(T_2,T_2) & \cdots & s(T_2,T_t) \\ \vdots & \vdots & & & \vdots \\ T_t & s(T_t,T_1) & s(T_t,T_2) & \cdots & s(T_t,T_t) \end{array}$$

That is, term $\bar{t}_k$ of the centroid is defined as the average value of all the values of TERM_k in the individual documents of the class, or $\bar{t}_k = \frac{1}{m}\sum_{i=1}^{m} t_{ik}$ for a class which has m term vectors. The term class refinement now consists in computing the similarity between each term vector TERM_k and each class centroid TERM-CENTROID for all existing classes. Assuming t term vectors and p classes, the process requires the generation of t × p similarity coefficients $\text{SIMILAR}(\text{TERM}_k, \text{TERM-CENTROID}_h)$ for k ranging from 1 to t, and h ranging from 1 to p. Each term vector is now entered into the class for which the similarity to the TERM-CENTROID is largest. If this involves a switch of a given term vector from one class to another, the centroids of those classes must be recomputed. This process can be pursued until no further class changes occur for the vectors, or until the number of class changes which occur after processing all the term vectors is sufficiently small [31].

Methods also exist for constructing a hierarchical arrangement of term classes, for example, by first building small classes consisting of a few terms exhibiting substantial pairwise similarities, and then expanding these initial classes into large groups. The new term groups subsume the initial classes by adding new terms whose similarity with the other terms already included in the class is successively weaker. Alternatively, large heterogeneous classes can be broken down into small more homogeneous entities by removing terms that have relatively weak similarities with the remainder.

Since the number of terms in a system is normally much larger than the number of thesaurus classes, thesaurus construction methods such as the single-link and clique methods which depend on the availability of all pairwise term similarities may be expensive to implement. Methods based on initial cluster assignments of the terms require less computer time. The available evidence indicates that thesauruses and automatically constructed term associations are quite effective in improving the recall performance provided the items entered into common classes exhibit high similarities. That is, $\text{SIMILAR}(\text{TERM}_k, \text{TERM}_h)$ should be large if TERM_k and TERM_h are entered into a common class. The high-frequency terms must be excluded from the thesaurus.

C Thesaurus Use

A thesaurus can be used to broaden the existing indexing vocabulary by replacing the initial terms with the corresponding thesaurus class identifiers, or by adding the thesaurus class identifiers to the original terms. A simple term expansion process which requires only the availability of term associations, but not of formal thesaurus classes, is illustrated in the example of Fig. 3-4. Assuming one already has term similarity information such as that provided by a term-term similarity matrix (see Table 3-11), a threshold K is chosen. This is used to transform the original term-term matrix into a binary form by replacing each

$$\begin{array}{c|ccccc} & A & B & C & D & E \\ \hline A & 1 & 1 & 0 & 0 & 0 \\ B & 1 & 1 & 0 & 1 & 0 \\ C & 0 & 0 & 1 & 0 & 1 \\ D & 0 & 1 & 0 & 1 & 1 \\ E & 0 & 0 & 1 & 1 & 1 \end{array}$$

(a)

Original term	Associated terms
A	B
B	A, D
C	E
D	B, E
E	C, D

(b)

$$q = \begin{pmatrix} A = 4 \\ B = 2 \\ C = 1 \\ D = 1 \\ E = 0 \end{pmatrix} \quad \begin{array}{l} \text{add } B = 2 \\ \text{add } A = 1,\ D = 1 \\ \text{add } E = \tfrac{1}{2} \\ \text{add } B = \tfrac{1}{2},\ E = \tfrac{1}{2} \\ \text{add nothing} \end{array} \quad q' = \begin{pmatrix} A = 5 \\ B = 4\tfrac{1}{2} \\ C = 1 \\ D = 2 \\ E = 1 \end{pmatrix}$$

(c)

Figure 3-4 Sample process for utilization of term associations. (a) Sample binary term-term similarity matrix for five terms (A through E). (b) Corresponding term associations. (c) Alternative associative indexing strategy. (Add associated terms with weight equal to one-half the original.)

matrix element by 1 whenever the value in the term similarity matrix is greater or equal to K and by 0 when the value is less than K.

A sample binary term-term similarity matrix for five terms, labeled A through E, is shown in Fig. 3-4a. The corresponding term association information is detailed in Fig. 3-4b. Given a particular term vector, such as that labeled q in Fig. 3-4c, it is now possible to add the information about the associated terms with a weighting factor (arbitrarily selected here as one-half of the original weight in the illustration of Fig. 3-4c). It may be noted that in the example one new term has been added to the original vector (E) since its weight is now greater than 0. The weight of several other already existing terms (A, B, and D) has also been altered [32–34].

Term associations and thesaurus classes can be displayed in a variety of formats to help the information system users in formulating the search requests and familiarizing themselves with the vocabulary. One attractive format is based on a graphlike structure where the nodes represent the individual terms and different kinds of lines denote different strengths of association [different magnitudes of the similarity coefficient SIMILAR($TERM_k$,$TERM_h$)] between the terms. An excerpt from such a term association map is shown in Fig. 3-5 [35].

One major disadvantage inherent in the use of any thesaurus is the necessity to maintain it. Two different maintenance problems arise. First, the thesaurus may require rebuilding as a result of user interaction with the system. For instance, new queries may be submitted for which the current thesaurus is inadequate, or new user populations and interests may appear which in turn require new vocabulary terms. Second, a thesaurus maintenance system may be needed to accommodate collection growth. When new documents are added to a collection, several updating strategies are possible:

1 The original thesaurus might be left unchanged and used for the expanded collection.

2 New terms derived from the added items might be placed into existing thesaurus categories only.

3 New terms might be placed into separate new classes.

4 The thesaurus might be completely restructured by generating a term classification from the updated vocabulary.

The fourth alternative may be very expensive. Hence it is necessary to consider one of the other possibilities. The available evidence indicates that some performance loss is produced when a thesaurus constructed for an original document environment is later used for an updated collection [36]. Unfortu-

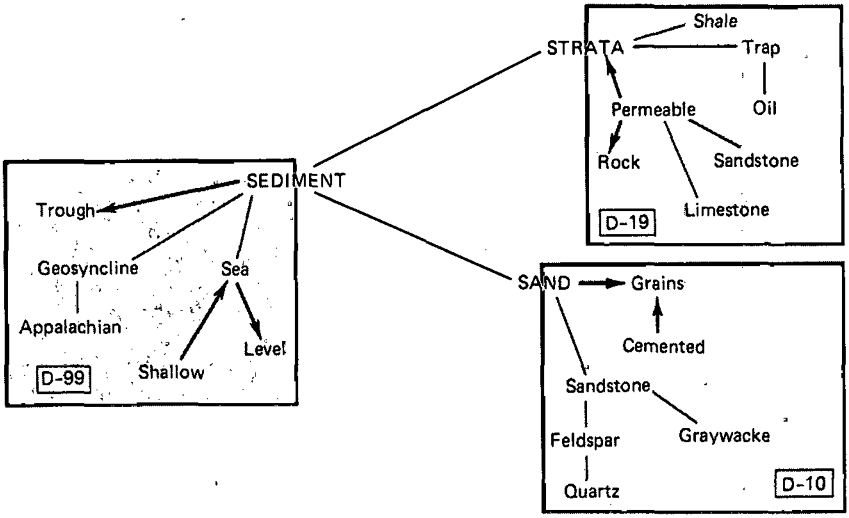

Figure 3-5 Term association map. (Different types of connecting lines denote different strengths of associations between terms). (*Adapted from reference 35.*)

nately, no experimental data are available leading to a clear-cut choice between the second and third alternatives.

Before leaving the subject of automatically generated term associations, it should be mentioned that recall-enhancing expansions of the vocabulary can be generated by using identifiers that are not standard index terms. Specifically, documents exhibiting similarities in bibliographic citation patterns—either references included in the reference lists attached to the documents or citations to the various documents made by other documents—may also reveal similarities in subject content [37,38]. This finding suggests that improved retrieval may be obtained by adding this citation information to documents and query vectors in addition to the normal subject identifiers. Alternatively, the standard subject terms could be replaced by the citation patterns.

One possibility is to lengthen the document vectors by including bibliographic reference indicators to documents outside the collection. Search requests (query vectors) can then be similarly lengthened by adding identifiers of relevant documents designated by the users. One system of that kind is sketched out in Fig. 3-6. The available experimental evidence indicates that substantially better retrieval results are obtainable with the augmented vectors including citations than with standard vectors consisting of subject indicators only. To utilize such a system, it is however necessary to store substantial citation information for the collection, and to obtain information regarding the relevance of documents from the users of the system. The use of bibliographic citations is examined in more detail in Chapter 6.

D Construction of Term Phrases

The recall of a search may be improved by broadening the terms used in query and document specifications and by adding new associated terms. On the other

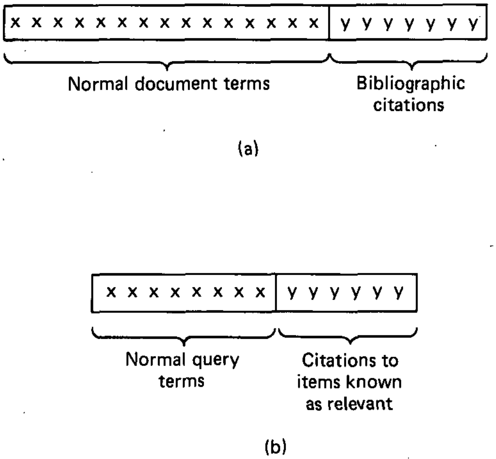

Figure 3-6 Expanded document and query vectors. (a) Expanded document vector. (Each x represents a standard index term, each y a bibliographic reference.) (b) Expanded query vector.

hand, the precision may be improved by using specific terms or by using terms in combination with each other. Terms of high specificity can be identified simply by their concentration in a few documents of a collection, as described earlier. To generate combinations of terms (phrases) one uses two or more terms, say TERM_k and TERM_h, with particular occurrence properties, and replaces them by a phrase (PHRASE_{kh}). For example, "computer" and "program" may be replaced by "computer program" or "computer programming." One expects the frequency of occurrence of PHRASE_{kh} in the document collection to be smaller than that of TERM_k or TERM_h; furthermore, the phrase term will have a more specific interpretation than the individual phrase components.

Various phrase-generation methods are possible, including the use of syntactic language analysis. These methods are used to identify phrases whose components exhibit acceptable syntactic relationships. However, consider first a phrase-generation process based on frequency considerations like those used earlier to generate individual terms. The best phrases (word pairs, triples, etc.) may include terms whose joint frequency of occurrence in the collection is larger than expected, given the frequencies of the individual terms. If PAIR-FREQ_{kh} is the total pair frequency in the collection of TERM_k and TERM_h, and TOTFREQ_k and TOTFREQ_h represent the collection frequencies of the individual terms, then the *cohesion* of the term pair may be defined as

$$\text{COHESION}_{kh} = \text{SIZE-FACTOR} \cdot \frac{\text{PAIR-FREQ}_{kh}}{\text{TOTFREQ}_k \cdot \text{TOTFREQ}_h} \tag{13}$$

SIZE-FACTOR represents a factor related to the size of the indexing vocabulary. Phrases can now be chosen as term pairs with a sufficiently high cohesion factor, subject to certain restrictions.

To utilize the preceding formula, it is necessary to choose an appropriate context for determining when two or more terms co-occur. In principle, it is possible to choose a wide context by declaring that two terms co-occur whenever they are included in a common document. Better (higher-precision) results may be obtainable by restricting the context to terms occurring in the same sentences of particular documents, or in the same sentences but with at most k words occurring between components, or in the same sentences in adjacent word positions, or finally in the same sentences in adjacent word positions and in the correct word order. When the context used to define a phrase is restricted, the phrase detection process becomes more costly because tests must be made to ensure that the various restrictions are obeyed, and this in turn implies that word location information must be stored specifying the positions of the individual words in the sentences. Furthermore, when restrictions are imposed on the phrase formation process, the number of phrases generated for a given item will of course decrease.

In addition to using context for phrase definition purposes, it may also be important to invoke frequency restrictions on one or more of the components of each phrase. As has been stated earlier, the phrase formation process is designed to create specific content identifiers to enhance precision. When rare terms that are already specific in the documents are combined into phrases, the resulting phrase terms may turn out to be overspecific and the retrieval results may deteriorate. Therefore, the phrase formation process should be restricted to include only relatively broad, high-frequency components.

In summary, a reasonable phrase formation method based on statistical word occurrence properties would define a phrase from word pairs with sufficiently high cohesion factors. The phrase components must occur in the same sentences of the document, and at least one component in each phrase should have a document frequency exceeding a stated threshold.

Statistical procedures for the generation of phrase identifiers are often assumed to be unreliable because they lead to the identification of statistically meaningful but syntactically incorrect phrases. Examples show that the statistical methodology leads to unfortunate results, such as, for example, the confusion of "blind Venetian" and "Venetian blind." This particular objection may however not be serious, since among other things these phrases are unlikely to occur in the same documents. Nevertheless, it is important to determine whether more sophisticated linguistic tools might be used to control the automatic indexing process.

The approaches using syntactic and/or semantic analysis features have not met with much success. This is largely because of the technical inadequacy and the excessive cost of the linguistic procedures. However, some advances have taken place in these areas which are mentioned briefly in the next section of this chapter and are covered in more detail in Chapter 7.

It is possible now to propose an automatic indexing process based on simple, well-understood procedures, capable of producing high-performance retrieval results. The word stems occurring in document titles and abstracts are isolated and term weights are computed using either inverse document frequencies [expression (2)] or term discrimination values [expression (9)]. Three classes of terms are then identified. Those in the middle-frequency ranges with positive discrimination values, or frequency characteristics, are used as index terms directly without further transformation. The broad high-frequency terms with negative discrimination values and excessive document frequencies are either discarded or incorporated into phrases with lower-frequency characteristics. Finally the narrow low-frequency terms with discrimination values close to zero are broadened by inclusion into thesaurus categories. The thesaurus class identifiers are then used as index terms for content representation [22,32].

The process is represented schematically in Fig. 3-7 where a document frequency axis is used to arrange the terms into three classes. The terms in the center are used unchanged; the broad terms are subjected to the phrase-formation process, represented by a right-to-left transformation in Fig. 3-7; the narrow, specific terms are incorporated into term thesaurus classes and are represented by the corresponding left-to-right transformation of Fig. 3-7.

The automatic process will produce a large number of content identifiers for each item—typically a hundred terms or more may be automatically as-

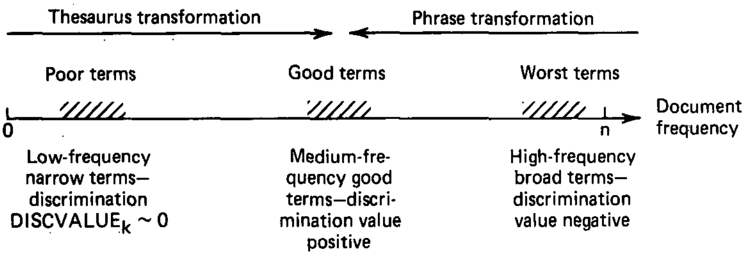
Thesaurus transformation
Phrase transformation
Poor terms
Good terms
Worst terms
0
n
Document frequency
Low-frequency narrow terms—discrimination DISCVALUE$_k$ ~ 0
Medium-frequency good terms—discrimination value positive
High-frequency broad terms—discrimination value negative

Figure 3-7 Term characterization in frequency spectrum.

signed compared with the half dozen terms used in a manual system. This fact explains in part the high order of performance of automatic indexing systems. Some laboratory evaluation data for automatic indexing strategies are presented in the last section of this chapter.

E Automatic Sentence Extraction

Text processing methods based on a determination of term or sentence importance have been used not only for indexing but also for automatic abstracting purposes. Ideally, given a document represented as natural language text, one would like to construct a coherent, well-written abstract that informs the readers of the contents of the original, or at least indicates whether the full version may be of interest to the reader. In fact, most procedures carry out an extraction process in which the abstract is defined simply as a small set of sentences pulled from the original, which are deemed to be important for purposes of content representation.

The extracting methods used over the years all start with a calculation of word and sentence significance, similar in spirit to the computation of the term weights in automatic indexing [12,39,40]. Criteria for the selection of important terms may be *positional*, involving, the place in the document where a particular term is located (for example, in the summary, title, etc.); they may be *semantic*, involving for example the relationship of this word or sentence to certain other words; or they may be *pragmatic*, such as a system which would consider proper names as highly significant. Furthermore, *statistical weights* based on term frequency or term distribution characteristics may be used in addition to the above criteria.

Given an indication of term significance, it is possible to define the importance of a phrase as a function of the weight of the individual terms and of the distance between the significant phrase components (the number of words between them). Thus if two terms have weights WEIGHT_i and WEIGHT_j, respectively, a phrase developed by combining the two terms might be assigned a weight equal to

$$\text{PHRASE-WEIGHT} = \frac{1}{2^{\text{DISTANCE}}} (\text{WEIGHT}_i \cdot \text{WEIGHT}_j) \tag{14}$$

where DISTANCE equals the number of intervening words. By extension, a significant sentence may be one which contains a large number of significant word groups. A flowchart for a typical sentence extracting process is shown in Fig. 3-8.

Because pure frequency characteristics are not likely to be reliable for either indexing or extracting, a variety of additional criteria have been used experimentally in an effort to obtain more satisfactory extracts. In particular, the word and sentence contexts can be taken into account in determining the *contextual inference* and *syntactic coherence* criteria [41]. Contextual inference means that the context within which a given word or phrase is placed in a document is used in addition to other criteria in an effort to decide on sentence se-

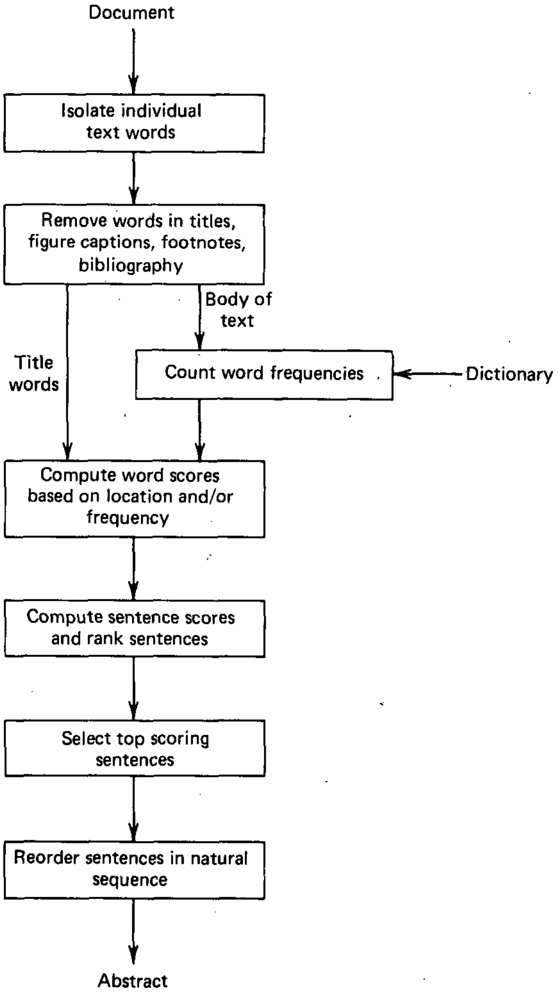
Document
Isolate individual text words
Remove words in titles, figure captions, footnotes, bibliography
Title words
Body of text
Count word frequencies
Dictionary
Compute word scores based on location and/or frequency
Compute sentence scores and rank sentences
Select top scoring sentences
Reorder sentences in natural sequence
Abstract

Figure 3-8 Typical sentence extracting system. (*Adapted from reference 39.*)

lection or rejection. Contextual inferences may be based on *sentence or word location* or on the presence of so-called *cue words*. Thus, sentences occurring under certain headings may be particularly important; the same may be true of sentences occurring very early or very late in the paragraph structure. Similarly sentences ending with a question mark are normally rejected, as are certain sentence portions occurring between pairs of commas.

The cue method is based on the presence of positive or negative indicators of sentence value. Thus, the presence of phrases such as "our work," "this paper," and "the present research" is assumed to introduce a statement that should be included in an abstract. Contrariwise, opinions and references to figures or tables that might be identified by "obvious," "believe," "fig. 1," etc., lead to sentence rejection. To operate effectively, the cue method depends on the availability of a dictionary containing the cue words together with indications of their semantic and/or syntactic value. Such a cue word dictionary may be particularly valuable for sentence rejection, as opposed to selection, because rejection can often be based on the presence of a small number of frequently occurring words, whereas selection may depend on longer lists of desirable words.

No matter how sophisticated the extraction process, a set of extracted sentences is not likely to constitute a coherent whole. Even if the extract in fact consists of appropriate topic sentences, the flow of ideas from one sentence to the next is likely to be interrupted because the discourse and reasoning leading to, or following from, the selected sentences is probably absent.

Coherence criteria based on syntactic or semantic considerations can be used to mitigate to some extent the shortcomings inherent in an automatic extracting process. Thus words or phrases indicating intersentence reference can be included in the cue dictionary ("these," "they," "it," "above," "presented earlier," "stated above," etc.), and their presence in a given extract can be used as a clue for inclusion in the abstract of the earlier sentences being referred to. Similarly, if the same important terms occur in adjacent sentences, a presumption exists that the sentences are related, and both should probably be included or excluded [41].

No matter what is done, a stylistically beautiful abstract is not likely to be created automatically because the linguistic difficulties are simply too severe for an effective automatic treatment. For example, the clarification of ambiguous antecedents such as those of certain pronouns is notoriously difficult. Some of the linguistic problems are further discussed in the next section together with other theoretical questions.

The overall conclusion is that automatic abstracting is less developed than automatic indexing and less likely to be used on a production basis in the near future. Abstracts must be placed in a readable natural language context and must obey the normal stylistic constraints. Sets of index terms, on the other hand, are not burdened by stylistic rules. Readable extracts are obtainable without excessive difficulties, but perfection cannot be expected within the foreseeable future.

6 SOME THEORETICAL APPROACHES

*A The Use of Linguistic Methods

It has been mentioned that the absence of syntactic recognition features may cause problems in the construction of indexing phrases capable of reflecting correctly document or query content. Two phrase construction problems, in particular, may be solved by using linguistic tools: first, the coordination of terms in accordance with the available context, and second, the assignment of roles to the phrase components. Given an indexing description such as "hardness, density, titanium, water," it is not clear a priori which qualification (hardness or density) applies to which material (titanium or water); and in "blind, Venetian," blind could presumably function as a qualifying adjective, or alternatively as the governing noun.

The approaches using linguistic methods in information retrieval are really of two kinds: on the one hand, it is possible to use simple methodologies with limited aims such as removing the ambiguity from some noun phrase identifiers; on the other hand, more complex linguistic analysis systems can be utilized but the context in which these systems operate must be limited. In the present discussion, it is possible only to give a very brief introduction; more details are included in Chapter 7 of this volume. A variety of other sources exist describing the use of linguistic methods in retrieval [42–44].

Consider the use of simple syntactic aids for the construction of indexing phrases. Normally, a *context-free phrase structure grammar* is used to obtain for each document, or query sentence, a *parse tree* which shows the syntactic structure of the sentence. A context-free grammar decomposes a sentence into nested and juxtaposed sentence portions, as in the example of Fig. 3-9, where "the man" and "the ball" are identified as noun phrases and "hit the ball" as a verb phrase. Simple phrase structure grammars can be used to recognize many types of noun phrases and prepositional phrases that might constitute useful document identifiers.

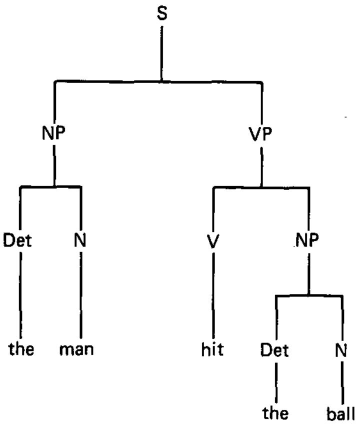

Figure 3-9 Sample context-free phrase structure analysis (S = sentence; NP = noun phrase; VP = verb phrase; Det = determiner; N = noun; V = verb).

Unfortunately, a simplified language analysis system based on context-free grammars exhibits a number of disadvantages. First of all, some sentences whose structure is not of the basic phrase structure type cannot be analyzed by the phrase structure model. Second, a unique analysis pattern is not obtainable for many sentences, but multiple parse trees can be generated—all ostensibly correct according to the grammar in use—without any information that would identify which of those analyses may be semantically acceptable. The notorious sentence "time flies like an arrow" may serve to illustrate a case where at least four reasonable analyses are generated by a typical context-free analysis system [45]:

1 Time passes as quickly as an arrow flies.
2 You should time the flies as quickly as an arrow times the flies.
3 You should time the flies which are similar to an arrow.
4 There exists a species of flies, called "time flies," which are fond of an arrow.

Most important of all, the phrase structure model is not sufficiently rich to make it possible to recognize semantic relationships between sentence components that may not be reflected by some sort of physical juxtaposition of the components in the sentence. Thus a requirement that the phrase components be grammatically related according to the phrase structure model—for example, that the components appear in the same "subtree" within the parse tree—may actually amount to an overspecification, producing an underassignment of phrases for the respective documents. Consider as an example the sentence "people in need of information require effective retrieval systems." In this sample sentence the terms "information" and "retrieval" are not related in the usual noun phrase sense, where "information" is grammatically dependent and modifies the governing element "retrieval." Hence if the indexing rule consisted in requiring a noun phrase or prepositional phrase relationship between phrase components, the phrase "information retrieval" would not be assigned to a document containing this sentence. On the other hand, a frequency-based phrase assignment method of the type covered earlier would correctly generate the phrase "information retrieval."

Various attempts have been made to use simple syntactic analysis systems in actual information retrieval situations. While linguistic methods may eventually prove essential in automatic indexing, the available evidence indicates that the simplified syntactic analysis systems do not yet provide the answer. The frequency-based phrase-generation methods are simpler to implement and are currently more effective [46–50].

At least two possibilities are apparent for helping with the syntactic analysis problems. On the one hand, certain *computer-aided* indexing systems have been proposed where the indexers obtain access to a computer console during the indexing process. In such circumstances, the human operators can intervene at appropriate times during the indexing process. This can be done, for example, by choosing the correct analysis output from among a number of parse trees or by identifying "good" phrases from among a large selection of potential phrases produced by statistical or syntactic procedures [51].

Alternatively, sophisticated linguistic methods are available for carrying out the text analysis. Specifically, syntactic models can be used in which contextual constraints exercised at each level of the analysis are used to restrict the number of syntactic parses [52,53]. These constraints, or restrictions, may be subject matter–specific, or they may depend on the inclusion of semantic data such as "case frames" which identify the (semantic) role of the sentence components.

A conventional parse tree represents the "surface" structure of a sentence, and the results of an initial parsing operation must be subjected to certain transformations before the underlying "deep" (that is, semantic) structure is obtained. The conventional (surface) parse for "John is easy to please" recognizes "John" as the subject of the sentence and "easy to please" as the complement. The deep structure, on the other hand, reflects the meaning of "It is easy for someone (unmentioned) to please John." In that case "John" appears properly as the complement.

Unfortunately the construction of transformational grammars capable of performing the required transformation operations has proved to be too difficult for practical purposes. Indeed, in conventional situations, dozens of possible transformations could in principle be applied at each point to a given surface structure, and no guidance is available for choosing the correct pattern of transformations. Here again substantial progress is evident in recent years, based on the use of restricted areas of discourse and of sufficient context to produce for each sentence only a small number of surface structures, each being subject to only a small number of possible transformations.

It now appears that when the input material is restricted to certain specialized topic areas, limited vocabularies, and limited syntactic patterns, *canonical representations* are obtainable from natural language input. The canonical representations consist of standard forms in which each phrase in the text is assigned a well-defined role, reflecting the full complexity of the syntactic and semantic structures of the text. The canonical forms for the documents can be compared with standard query forms to derive answers to incoming user queries. For the most part, linguistic procedures are incorporated into specialized *question-answering* systems where direct responses are given in answer to search requests, as opposed to answers consisting of document references that must be consulted before direct answers can be obtained [54–58].

In at least two cases, the natural language text of full-length medical records, such as radiology reports and pathology diagnostic reports, has been transformed automatically into standard tabular formats where each text component is identified as to function and meaning. Such tabular formats can then be directly transformed into structured data bases used to generate answers to search requests [59–60]. Such methods may eventually be used in unrestricted automatic indexing environments; however, the linguistic procedures must then possess substantial sophistication, and their application will probably be very costly. It remains to be seen whether this route will actually be pursued in the future.

*B Fragment Encoding

Indexing methods, based on probabilistic considerations, like the previously discussed linguistic methods, offer great promise, although really practical methods applicable on a wide scale have not so far been generated. One line of effort, somewhat tangential to the main developments, is based on the previously mentioned information theoretic considerations relating to the uneven nature of the occurrence frequencies of the standard indexing products. Not only is the number of index terms needed to represent the content of most document collections extremely large—even small collections of a few hundred documents may require several thousand different terms—but many of these terms occur very rarely. Hence their utilization, storage, and maintenance is quite inefficient.

In these circumstances, it is not surprising that the suggestion has been made to replace the normal index words or terms by a small number of artificial entities exhibiting approximately equal occurrence frequencies in the documents of a collection. Specifically, variable length character strings known as "fragments" can be used for indexing instead of full-length words. Normally fragments represent substrings of complete words, or terms. A redundant encoding is often used so that certain characters in the original terms are repeated in several different fragments. For search purposes, a number of different fragments must then be used in a given document or query to replace each original term [61–66].

A variety of encoding procedures can be used to construct a fragment set consisting of more or less equally occurring elements. Consider the following procedure [66]:

1 A string of n characters is created starting from *each* character of the original text sample; n must be chosen at least as large as the longest fragment expected to occur in the final fragment set.

2 The character strings are sorted and the frequency of each string is determined.

3 The frequency of each distinct character string is then compared with a given threshold frequency (the size of the final fragment set is inversely related to the magnitude of the threshold).

4 Any string whose occurrence frequency exceeds the threshold is selected for inclusion in the final fragment set, and eliminated from the set of strings still under consideration.

5 Strings whose frequency does not exceed the threshold frequency are shortened by truncation of the rightmost character, the equivalent shortened strings being merged; new frequencies are computed for the shortened strings.

6 Any shortened character string whose frequency now exceeds the threshold is selected for inclusion in the final fragment set, and the procedure is repeated until single character strings are reached.

7 The final fragment set also includes all single character strings.

The original set of index terms must now be mapped into a final fragment set. Various methods suggest themselves such as, for example, the "longest match" algorithm which chooses the longest fragment that matches the beginning (left-hand end) of each term. The longest fragments matching the left end of the remainder of the various words are then successively chosen until the whole text is covered [67]. For efficiency in the final encoding, it is convenient to operate with a total number of fragments equal approximately to a power of 2. Thus, each of 256 fragments can conveniently be represented by an 8-bit code. The fragment encoding for some typical terms is illustrated in Table 3-12.

A thesaurus, or inverted index, giving access to a fragment encoded document set is managed more easily than a standard index, because a dictionary of a few hundred fragments can replace the normal list of many thousands of conventional index terms. However, a "false drop" problem must be contended with in retrieval because certain sets of fragments included in a search request may correspond to several different full index terms. Thus the two fragments RAC and ER can correspond to RAC/ER and also to T/ER/RAC/E, causing documents on terraces to be retrieved when racers are wanted. The seriousness of this problem depends on the actual keyword set used for a particular collection.

A great many automatic text compression methods other than fragment encoding are in use, designed to take advantage of the unequal occurrence probabilities of individual characters and words, and of the redundancies built into the natural language [68]. A description of these methods is beyond the scope of the present discussion.

**C Probabilistic Information Retrieval

Probability theory has also been used as a principal means for modeling the retrieval process in mathematical terms. In conventional retrieval situations, a

Table 3-12 Sample Fragment Encoding

Original term	Fragments
AMINO ACID	AMINO, ACID
BETA HYDROXYLASE	HYDRO, BETA, OXY, YLA, ASE
BETA-HYDROXY LASE	BETA-, -HYDR, LASE, ROX, XY
DICHLOR ACETAT	CHLOR, ACETA, DIC, TAT
HALOGEN	HALO, OGEN
CHLOROFORM	CHLOR, FORM, RO, OF
SULFONAMID	AMID, SUL, FON, LF, NA
IONISATION	ATION, IONI, ISA

Adapted from reference 61.

document is retrieved in response to a query whenever the keyword set attached to the document appears similar in some sense to the query keywords. In this case the document is assumed to be relevant to the corresponding query. More explicitly, since relevance of a document with respect to a query is a matter of degree, one postulates that when the document and query vectors are sufficiently similar, the corresponding probability of relevance is large enough to make it reasonable to retrieve the document in answer to the query.

More formally, the retrieval problem can be expressed as a decision-theoretic process using three basic parameters as follows: P(Rel), the probability of relevance of a record; LOSS_1, a loss parameter associated with the retrieval of a nonrelevant or extraneous record; and LOSS_2, a loss associated with the nonretrieval of a relevant record. A loss minimizing rule can be devised by noting that the retrieval of an extraneous item causes a loss of $[1 - \mathrm{P(Rel)}] \cdot \mathrm{LOSS}_1$, whereas the rejection of a relevant item produces a loss of $\mathrm{P(Rel)} \cdot \mathrm{LOSS}_2$. In these circumstances the total loss is minimized by opting for retrieval of an item whenever

$$\mathrm{P(Rel)} \cdot \mathrm{LOSS}_2 \geq [1 - \mathrm{P(Rel)}] \cdot \mathrm{LOSS}_1 \qquad (15)$$

Equivalently a discriminant function DISC may be defined, and an item may be retrieved whenever DISC ≥ 0 [69–71], where

$$\mathrm{DISC} = \frac{\mathrm{P(Rel)}}{1 - \mathrm{P(Rel)}} - \frac{\mathrm{LOSS}_1}{\mathrm{LOSS}_2} \qquad (16)$$

A retrieval rule of the kind produced by equation (16) is not useful in practice because the relevance properties of the individual records cannot be divorced from other system parameters. Thus, it becomes necessary to relate the discriminant function to other design parameters, and most notably to the *indexing* process. This can be done by defining two conditional probability parameters:

$\mathrm{P(TERM_i \mid Rel)}$ = the probability of $\mathrm{TERM_i}$ occurring in a document given that the document is relevant to a given query

and

$\mathrm{P(TERM_i \mid Notrel)}$ = the probability of $\mathrm{TERM_i}$ occurring given that the document is not relevant to the query [71,72].

Using a formula developed by Bayes, a retrieval function $\mathrm{P(Rel \mid Doc)}$ can be obtained, representing the probability of relevance given a document Doc = $\langle \mathrm{TERM}_1, \mathrm{TERM}_2, \ldots, \mathrm{TERM}_t \rangle$. In particular,

$$\mathrm{P(Rel|Doc)} = \frac{\mathrm{P(Doc|Rel)\ P(Rel)}}{\mathrm{P(Doc)}}$$

and (17)

$$\mathrm{P(Notrel|Doc)} = \frac{\mathrm{P(Doc|Notrel)P(Notrel)}}{\mathrm{P(Doc)}}$$

where P(Rel) and P(Notrel) are the a priori probabilities of relevance and nonrelevance of an item, and

$$\mathrm{P(Doc)} = \mathrm{P(Doc\,|\,Rel)} \cdot \mathrm{P(Rel)} + \mathrm{P(Doc\,|\,Notrel)} \cdot \mathrm{P(Notrel)} \tag{18}$$

If one assumes that the two loss parameters are equal to 1 ($\mathrm{LOSS}_1 = \mathrm{LOSS}_2 = 1$), then the obvious retrieval rule calls for retrieval whenever

$$\mathrm{P(Rel\,|\,Doc)} \geq \mathrm{P(Notrel\,|\,Doc)} \tag{19}$$

or whenever the discriminant function DISC ≥ 1 where

$$\mathrm{DISC} = \frac{\mathrm{P(Rel|Doc)}}{\mathrm{P(Notrel|Doc)}} = \frac{\mathrm{P(Doc|Rel)} \cdot \mathrm{P(Rel)}}{\mathrm{P(Doc|Notrel)} \cdot \mathrm{P(Notrel)}} \tag{20}$$

The foregoing rule relates the retrieval of the records to the occurrence characteristics of the terms in both the relevant and the nonrelevant items. For practical application, it is necessary to specify how the probabilities P(Doc | Rel) and P(Doc | Notrel) are to be determined. The problem is twofold in that one must first determine the occurrence characteristics for each term separately, and next the interactions between terms. In most abstract retrieval models the second problem is settled either by considering single-term queries only, where term interactions are of no consequence [71,73], or more drastically by disregarding term interactions altogether, and assuming that terms occur independently of each other in the records of the collection. The first question relating to the individual term occurrences can be handled either by using a probability distribution, such as the Poisson distribution, to characterize the occurrence characteristics of the terms, or by studying the actual occurrences of the terms in a typical sample record collection and applying the findings to other collections at large. (The Poisson distribution is an approximation to the binomial distribution which measures the probability of success in a sequence of success-failure experiments when the number of experiments is large and the probability of success of a given event is small. In the case at hand, the texts under consideration are assumed to be long, the word occurrences are randomly placed in the text, and the probability of occurrence of a particular word is assumed to be small.)

Consider the case where *term independence* is assumed and where the occurrence characteristics are obtained from a sample collection. In such circumstances one can write

$$P(\mathrm{Doc}\,|\,\mathrm{Rel}) = P(\mathrm{TERM}_1\,|\,\mathrm{Rel}) \cdot P(\mathrm{TERM}_2\,|\,\mathrm{Rel}) \quad . \; . \; . \quad P(\mathrm{TERM}_t\,|\,\mathrm{Rel}) \qquad (21)$$

It remains to determine the probabilities of each term in both the relevant and the nonrelevant items in a collection. Consider a sample collection of N records and assume that R records out of N are relevant to a given query Q and N − R items are nonrelevant. The term occurrence characteristics for a specific TERM_i are listed in Table 3-13.

If one assumes that the term occurrences in the sample record collection of Table 3-13 are typical of the term occurrences at large, one can postulate that $P(\mathrm{TERM}_i\,|\,\mathrm{Rel})$ and $P(\mathrm{TERM}_i\,|\,\mathrm{Notrel})$ representing the probabilities that a given TERM_i occurs in a document, given that the document is respectively relevant and nonrelevant, is equal to

$$P(\mathrm{TERM}_i|\mathrm{Rel}) = \frac{r_i}{R}$$

and (22)

$$P(\mathrm{TERM}_i|\mathrm{Notrel}) = \frac{n_i - r_i}{N - R}$$

where r_i represents the number of relevant documents in which TERM_i occurs, and $(n_i - r_i)$ is the number of nonrelevant documents with TERM_i.

By inserting the expressions (21) and (22) into expression (20) for the dis-

Table 3-13 Occurrence Table for One Term
(r Relevant Documents Contain Term; n − r Nonrelevant Documents Contain Term; R Relevant Documents Exist in All with Respect to Some Query Q in a Collection of N Documents)

	Relevant	Nonrelevant	
Term present	r	n − r	n
Term absent	R − r	N − n − (R − r)	N − n
	R	N − R	N

criminant function, it is not hard to show that for each query term that matches a given document term, an appropriate term weight is given by

$$\text{TERMREL} = \frac{r/(R - r)}{(n - r)/[N - n - (R - r)]} \tag{23}$$

where r and n − r represent the number of relevant and nonrelevant documents in which the given term occurs. The expression TERMREL, known as the *term relevance,* represents the ratio of the proportion of relevant to the proportion of nonrelevant items in which the term occurs. TERMREL is a term weighting function for the term akin to other term weighting functions (inverse document frequency, term signal, and term discrimination value) previously introduced. It differs from these earlier weighting systems in that the relevant documents to a given query are used to compute TERMREL. Relevance information of particular documents with respect to particular queries may be obtainable in modern on-line information retrieval systems where users have direct access to the system during the course of the retrieval operations [74–76]. The term relevance weights may be most profitably incorporated into these on-line retrieval environments [77–79].

It can be shown that if the terms occur independently of each other, and binary (as opposed to weighted) terms are used to represent the documents, then the optimum query will have terms weighted according to the term relevance factor TERMREL. Furthermore, given a document $D = \langle x_1, x_2, \ldots, x_t\rangle$ and a query $Q = \langle q_1, q_2, \ldots, q_t\rangle$ the best matching function SIMILAR(D,Q) between them is the inner product $\sum_{k=1}^{t} x_k q_k$ previously introduced in expression (11).

Some of the restrictions of the decision-theory model can be removed, for example, by introducing term dependencies [72]. Of more immediate interest may be the utilization of the probabilistic model for automatic indexing purposes. For expository purposes, the indexing problem may be considered to be a classification problem where m subject classes are given, each described by a class term vector. The indexing (classification) task then consists in assigning each document to the class whose subject area most nearly reflects the document description. In indexing terms, the assignment of a document to a subject class may simply imply the use of the corresponding class identifiers to represent individual document content.

In the earlier development, the documents were characterized by the term occurrence probabilities in two classes, the class of relevant and nonrelevant documents, respectively, with respect to a given query. An alternative (classification) context represents a generalization of that used earlier, in that m classes $C_1, C_2, \ldots, C_m$ may be assumed to exist instead of only two. The decision rule now specifies that a document $D = \langle x_1, x_2, \ldots, x_t\rangle$ is assigned to the class for which its probability of occurrence is the largest.

A discriminant function can be constructed as before in terms of the probabilities of occurrence of the individual terms in the various classes. Eventually, each document is assigned to that class which exhibits the largest sum of all the probability values of the various document terms [80–82].

In the previous probabilistic models it was assumed that occurrence frequencies were available to characterize the term occurrences in the various document classes. Unfortunately, in practice, these frequencies are often difficult to generate. An alternative approach may then serve in which a document collection is broken down into homogeneous document classes such that the occurrence properties of the index terms may be characterized by an overall probability distribution with known parameters in each individual document class. In particular, if the terms are assumed to be randomly scattered across the documents within each of the homogeneous classes, the previously mentioned Poisson distribution accurately reflects the term occurrence characteristics.

This fact has been used in many of the early automatic indexing models by noting that the common function words which do not indicate document content exhibit the same occurrence properties in *all* the documents of a collection, and thus are characterized by a single Poisson distribution. The specialty words, on the other hand, that are reflective of document content tend to be clustered in a few documents and a single Poisson formula cannot be used to represent their properties across the documents of a collection. Instead the collection is broken down into subcollections, and the assumption is made that a *different* Poisson distribution applies to a given term in each subclass with different parameters. Several attempts have been made to predict the usefulness of index terms based on an analysis of the term occurrence characteristics in the documents of a collection, followed by a comparison with the Poisson model [71,73,83–86].

7 AUTOMATIC INDEXING EXPERIMENTS

Many observers consider the use of automatic indexing and text processing methods to be acceptable only if substantial advantages can be demonstrated for the automatic system compared with the conventional manual situation, in the form of lower costs, greater speed of operations, or more extensive collection coverage. Unfortunately, it is a fact that comparative tests of indexing effectiveness (and efficiency) must normally be carried out under controlled conditions, using test collections of relatively small size with small sets of test queries. Under these conditions it is not difficult to show that the retrieval effectiveness of many systems operating with automatically assigned content indicators is at least equivalent to that obtaining for manually operated systems. However, it is hazardous to extrapolate test results obtained in a laboratory environment to operational situations possibly involving hundreds of thousands of items. This is particularly true with respect to an evaluation of operating efficiency, as opposed to effectiveness, because the cost or speed of operations in the laboratory reveals little about costs or speeds in practical environments.

A detailed recital of evaluation results in automatic indexing is unlikely to prove conclusive; however, a brief summary of some of the early evaluation studies may provide a useful demonstration of the *relative* effectiveness of various automatic procedures, and an indication of possible future trends. Several different evaluation approaches are possible [87]:

1 Title word studies where an attempt is made to compare index entries derived from document titles with automatically generated terms

2 Studies involving the comparison of automatically generated and manually assigned term or class indicators

3 Retrieval experiments in which the manual or automatic methodologies are actually used in a retrieval environment and an attempt is made to assess the effectiveness in terms of recall and precision

The title word studies conducted for a variety of subject areas such as medicine [88], chemistry [89], or law [90] indicate that for a large proportion of the documents, varying from 60 to 80 percent, some portions of the title are usable directly for indexing purposes. Furthermore, the number of titles which are totally useless in automatic indexing appears to be small in most subject areas—of the order of 10 percent.

These results are reinforced by studies based on a direct comparison between automatically generated sets of terms and their manual equivalents. The coefficient of similarity between manual and automatic index sets may then be taken simply as the proportion of common term assignments (that is, terms present in both the manual and automatic sets) to distinct term assignments that is

$$Q = \frac{C}{A + M - C}$$

where Q is the similarity coefficient between indexing sets, A is the number of distinct terms derived automatically, M is the number of distinct manually assigned terms, and C is the number of common assignments. Various tests of this general type have been performed, and the consensus is that about 60 percent agreement between manually and automatically produced term sets is obtainable [91,92].

While retrieval results based on a direct comparison of content identifiers may be interesting, they prove relatively little in the end, since the terms themselves are not the issue, but rather the retrieval performance obtainable with them. For this reason, the evaluation of *retrieval* results produced by a variety of different indexing methodologies has received considerable attention over the last few years. The first comparison of a conventional, manual indexing system with an automatic text processing system appears to be the one performed by Swanson in the late 1950s [93]. In that instance the manual utilization of a conventional subject-heading index was compared with a system based on words and phrases automatically extracted from the document texts; in addition, a thesaurus was also used to modify the words extracted from the documents. The test results indicate that the average retrieval performance over 50 queries and 100 documents was superior for the system based on automatic text analysis. It may be worthwhile to quote from the conclusions of this early test [93]:

> The apparent superiority of machine-retrieval techniques over conventional retrieval . . . will become greater with subsequent experimentation as retrieval aids for text searching are improved . . . (because) no clear procedure is in evidence which will guarantee improvement of the conventional (manual) system. . . . Thus even though machines may never enjoy more than a partial success in library indexing . . . people appear even less promising.

These original test results were later confirmed by additional experiments in which, for the first time, natural language queries were used instead of manually constructed query formulations [94]. Furthermore, an evaluation methodology very similar to that used to the present day was utilized, in the sense that documents were retrieved in decreasing order of similarity to the queries, the similarity score for an article being computed by summing the weights of those words in the article which coincided with the query words. With such a ranked list of retrieved documents, it is possible to compute recall and precision values following the retrieval of each document (or each nth document) producing a sequence of recall-precision pairs which can be plotted as a curve giving recall against precision, or listed in a table containing average precision values at certain fixed recall points.

Many additional experiments have been performed over the years, designed to evaluate a variety of automatic indexing theories, and in some cases attempts were made to measure large numbers of index language variations [24, 95–97]. The Cranfield and SMART studies are perhaps among the best known. The Cranfield II experiments were designed to measure many linguistic ''devices'' that are potentially useful for the representation of document content, including synonym dictionaries, hierarchical subject classifications, phrase assignment methods, and others [96,97]. While all indexing tasks were performed manually by trained indexers, the indexing rules were carefully specified and carried out in such a way as to simulate a computer assignment. A collection of 1,400 documents in aerodynamics was used with 279 search requests prepared by aerodynamicists. Three main indexing languages were utilized:

1 The *single terms* were content words chosen from document texts.

2 *Controlled* terms were single terms modified by consulting a manually constructed subject authority list (thesaurus).

3 Simple *concepts* were phrases obtained by concatenation of single terms.

Each of these basic languages was used with a variety of recall-improving procedures (synonym dictionaries, concept associations, term hierarchies, etc.) and precision-improving methods (assignment of weights, specification of term relations).

Somewhat unexpectedly, the retrieval results obtained were apparently counterintuitive, in the sense that the simple uncontrolled indexing language involving single terms produced the best retrieval performance, while the controlled vocabulary and the phrases (simple concepts) furnished increasingly worse results. To quote from Cleverdon [96]:

> The seemingly inexplicable conclusion . . . is that the single term index languages are superior to any other type . . . the single terms appear to have been near the correct level of specificity; only to the relatively small extent of grouping true synonyms (using a synonym dictionary) and word forms (using a suffixing process to generate word stems) could any improvement in performance be obtained. . . .
>
> Of the controlled term index languages that using only the basic terms gave the best performance; as narrower, broader or related terms are brought in . . . the performance decreases. . . .
>
> The simple concept (phrase) index languages were overspecific. . . .

In other words, the conclusion is that, on the average, the simplest indexing procedures which identify a given document or query by a set of terms, weighted or unweighted, obtained from document or query texts are also the most effective. Only the use of synonym dictionaries exhibiting groups of related terms could produce improvements in retrieval performance.

Such a result, if verified in other test environments, is of interest for two reasons: first, the single term indexing process is easier to implement automatically than the more sophisticated, seemingly less effective alternatives; and, second, if the single terms could be shown to operate at the correct level of specificity for the average user query, an automatic indexing process might become competitive in both cost and effectiveness with a manual indexing method.

The verification of the Cranfield results was provided by the extensive evaluation work carried out for some years with the SMART system [98,99]. The experimental, automatic SMART document retrieval system uses a variety of automatic text analysis and indexing methods, including synonym dictionaries, hierarchical term arrangements, statistical and syntactic phrase-generation methods, and the like, to generate sets of weighted content identifiers useful for the retrieval process. Information is retrieved by using a composite vector matching method producing for each query-document pair a coefficient of similarity. A ranking is obtained for the stored items in decreasing order of the query-document similarity, and a variable number of documents can be retrieved as required by the individual requestors. The output is evaluated in terms of recall and precision by processing the same search requests against the same document collections several times, while making selected changes in the indexing procedures between runs as previously explained.

In one extensive set of tests, document collections were used in the fields of computer engineering, aerodynamics, and documentation. A few quotations from the published results may suffice for present purposes [24,87]:

> The order of merit is generally the same for all three collections. . . .
>
> The use of unweighted terms (with weights restricted to 1 for terms that are present, and 0 for those that are absent) is generally less effective than the use of weighted terms. . . .
>
> The use of document titles alone is always less effective for content analysis purposes than the use of full document abstracts. . . .
>
> The thesaurus process involving synonym recognition performs more effectively than the word stem extraction method where synonyms and other word relations are not recognized. . . .
>
> The thesaurus and statistical phrase methods (where phrases are formed by statistical association of terms) are substantially equivalent in overall system performance; other dictionaries, including term hierarchies and syntactic phrases, perform less well.

Thus, the principal conclusions reached by the Cranfield project are borne out by the SMART studies: that computer language devices are not substantially superior to single terms used as indexing devices, and that sophisticated analysis tools are less effective than had been expected. These conclusions are perhaps not surprising, given the fact that a retrieval system is designed to serve a large, sometimes heterogeneous user population. Since users may have different needs and aims, the search requests range from survey or tutorial type questions to very detailed analytical queries. In these circumstances, an excessively specific analysis may be too specialized for most users.

Furthermore, the evaluation process is based on a performance criterion averaged over many search requests. This implies that analysis methods whose overall performance is moderately successful are preferred over possibly more sophisticated procedures which may operate excellently for certain queries but far less well for others. In practice, it may turn out that for each query type, a specific sophisticated analysis will be optimal, whereas for the average query the simpler type of indexing is best.

It remains to apply the evaluation methodology to a comparison of automatic indexing with conventional, manual indexing methods currently used under operational conditions, and to the more advanced, automatic indexing theories based on the term discrimination values and frequency transformations described earlier in this chapter. To answer the first question, certain automatic indexing methods incorporated into the SMART system were utilized together with the exhaustive evaluation work of the operating MEDLARS retrieval system performed at the National Library of Medicine. The MEDLARS system is based on a manual analysis of documents and incoming search requests, and on an exhaustive search of a stored collection of several hundred thousand documents. In the conventional MEDLARS environment, an average recall of 0.577 and an average precision of 0.504 is reported for 300 test queries processed against the complete MEDLARS collection, implying that an average search manages to retrieve almost 60 percent of what is wanted, while only half the retrieved items are not relevant [100].

These results were compared with the automatic indexing methods by applying both types of procedures to a subcollection from the full MEDLARS environment and a subset of the original queries used in the original MEDLARS evaluation. The subcollection was constructed by using for each query one or more documents known to be relevant to the user's needs as entry points to the Science Citation Index (SCI). For each query, 15 new documents were then obtained from the SCI such that each new item cited the original known relevant document. Obviously, such a process produces a set of potentially relevant documents, independent of either of the retrieval systems.

The main retrieval results are summarized in Table 3-14 [101,102]. It can be seen that a deficiency of about 15 percent for the automatic indexing method using word stem extraction techniques (from document abstracts) is reduced to a deficiency of only about 5 percent using the previously mentioned discrimination values to delete negative discriminators. When a thesaurus is used to recognize synonymous and related terms, a small advantage is produced for the automatic indexing process. Other procedures, related to search strategy rather than to indexing, eventually generate an advantage for the automatic system of about 30 percent in recall and precision [102].

It should be noted that the original MEDLARS searches retrieved a total

Table 3-14. Comparison of MEDLARS Controlled Indexing with SMART Automatic Indexing
(Cutoff in SMART Query-Document Similarity Measure Set to Retrieve a Total of 127 Documents)

Analysis method	Cutoff determining number retrieved	Recall	Percent difference from MEDLARS	Precision	Percent difference from MEDLARS
MEDLARS (controlled indexing)	Boolean search (exact match)	0.3117		0.6110	
SMART (word stems with frequency weights)	Query-document similarity 0.2201	0.2622	−16	0.4901	−19
SMART (word stems with discrimination value weighting)	Query-document similarity 0.2109	0.2872	−8	0.5879	−4
SMART (thesaurus)	Query-document similarity 0.3720	0.3223	+4	0.6106	0

of only 127 documents over the 29 test queries, whereas the total number of items determined to be relevant to the query set was equal to 284 documents. Since the threshold in the query-document similarity coefficient used for SMART must be set to retrieve exactly as many documents as MEDLARS had obtained earlier (namely 127), the maximum recall obtainable by either system is 127/284 = 0.4471, obtained when all retrieved items are found to be relevant. This recall ceiling, imposed by the test conditions, accounts for the low recall values obtained by both systems. The precision ceiling (maximum precision obtainable) is of course always equal to 1. One may interpret the results of the foregoing tests as showing that both conventional and automatic indexing methods produce equally good (or equally poor) retrieval results.

The question now arises to what extent the more advanced automatic indexing theories can produce improvements in the performance over the simple frequency-based indexing methods. Consider first the performance of the more refined term weighting system [18–22]. Table 3-15 contains average precision figures for 10 recall points, ranging from 0.1 to 1.0, averaged over 24 user queries for a collection of 425 documents from the world affairs section of *Time* magazine. As previously explained, a recall-precision table may be obtained by choosing various retrieval thresholds, that is, by retrieving a variable number

Recall	Binary weights BIN_{ik}	Term frequency weights $FREQ_{ik}$	Binary with IDF weights $BIN_{ik}/DOCFREQ_k$	Term frequency with IDF $FREQ_{ik}/DOCFREQ_k$
0.1	0.8257	0.7496	0.8085	0.8536
0.2	0.7555	0.7071	0.7741	0.7901
0.3	0.6754	0.6710	0.7114	0.7568
0.4	0.6224	0.6452	0.6328	0.7503
0.5	0.5708	0.6351	0.6218	0.6783
0.6	0.5299	0.5866	0.5673	0.6243
0.7	0.4618	0.5413	0.5124	0.5823
0.8	0.4087	0.5004	0.4384	0.5643
0.9	0.2959	0.3865	0.3374	0.4426
1.0	0.2854	0.3721	0.3188	0.4170

(a)

A. Binary weights BIN_{ik} vs. B. Term frequency weights $FREQ_{ik}$		A. Binary with IDF vs. B. Term frequency with IDF		A. Term frequency $FREQ_{ik}$ vs. B. Term frequency with IDF $FREQ_{ik}/DOCFREQ_k$	
t-test	0.0000	t-test	0.0000	t-test	0.0000
B better than A		B better than A		B better than A	
Wilcoxon	0.0000	Wilcoxon	0.0000	Wilcoxon	0.0000

(b)

Table 3-15 Comparison of binary term weighting with inverse document frequency (IDF) weights (time collection, 425 documents, 24 search requests). (a) Comparison of binary and term frequency weighting with and without inverse document frequency normalization. (b) Statistical significance output for the results of Table 3-15a.

of documents in decreasing order of the query-document similarity, and computing a recall and a precision value for each retrieval threshold. The various recall-precision values can then be plotted on a graph or tabulated as shown in Table 3-15. Detailed methods for the construction of recall-precision tables and graphs are covered in Chapter 5.

It may be seen first of all that an unequivocal result is not obtained for the comparison between binary (BIN_{ik}) and term frequency ($FREQ_{ik}$) weights. In the binary case, the terms are not weighted, whereas a weight proportional to the term occurrence frequency in each document is assigned to each term in the other case. The preferred result is shown in each case by a vertical line in the second and third columns of Table 3-15a. For the collection under study, the binary weights are superior at the low recall end and the term frequency weights at the high recall end. When these basic term weights are combined with an inverse document frequency (IDF) factor proportional to $1/DOCFREQ_k$, the combined weighting system equivalent to $FREQ_{ik}/DOCFREQ_k$ [see expression (2)] is clearly superior. Indeed, the results in the rightmost column of Table 3-15a show improvements ranging from 3 percent at very low recall to over 30 percent at high recall.

Table 3-15b contains t-test and Wilcoxon signed rank test values, giving in each case the probability that the output results for the two runs being compared could have been generated from the *same* distribution of values. Small probabilities—for example, those less than 0.05—indicate that the answer to this question is negative and that the test results differ significantly (hence that one system is clearly superior to the other). The significance figures in Table 3-15b show that for the *Time* collection the method labeled B is significantly better than the A method in all cases. The significance tests used in judging evaluation output are examined in more detail in Chapter 5.

In Table 3-16 the two basic term weighting systems are combined with a term deletion process which eliminates from the document and query vectors those terms deemed to be poor content identifiers. A variety of term deletion procedures are usable in practice [103]. For test purposes two term deletion systems were used to obtain the results of Table 3-16. The first one consists in deleting terms in increasing IDF order, that is, terms exhibiting the highest document frequencies (which may be expected to produce the poorest index terms) are eliminated first. In the other case, terms were deleted in increasing term discrimination order by removing first the terms with the lowest discrimination values. The two runs are labeled IDF CUT and DISC CUT respectively in Table 3-16. For the *Time* collection under study, documents with a document frequency greater than 104 (out of 425 documents) were actually removed to generate the IDF CUT performance. For the DISC CUT performance, the term deletion was restricted to terms with negative discrimination values.

The placement of the vertical bars in Table 3-16 shows that the deletion in inverse document frequency order performs best at low recall, whereas the discrimination value cutoff is best at high recall. A comparison of the results of Tables 3-15 and 3-16 shows that the removal of poor terms produces better re-

Table 3-16 Recall-Precision Results for Term Deletion Methods
(*Time* 425 Collection, 24 Queries)

Recall	Standard BIN_{ik} weights	Standard FREQ_{ik} weights	FREQ_{ik} weights		A. IDF CUT vs. B. Standard FREQ_{ik}	A. DISC CUT vs. B. Standard FREQ_{ik}
			IDF CUT	DISC CUT		
0.1	0.8257	0.7496	0.8601	0.7911		
0.2	0.7555	0.7071	0.8268	0.7485	t-test	t-test
0.3	0.6754	0.6710	0.7503	0.7362	A better than B	A better than B
0.4	0.6224	0.6452	0.7144	0.7000	0.0000	0.0085
0.5	0.5708	0.6351	0.6872	0.6777		
0.6	0.5299	0.5866	0.6168	0.6350	Wilcoxon	Wilcoxon
0.7	0.4618	0.5413	0.5645	0.5907	A better than B	A better than B
0.8	0.4087	0.5004	0.5017	0.5510	0.0000	0.0127
0.9	0.2959	0.3865	0.4071	0.4177		
1.0	0.2854	0.3721	0.3906	0.4019		

IDF = inverse document frequency
DISC = discrimination value

sults than the composite inverse document frequency weights at low and medium recall points. The statistical significance probabilities on the right-hand side of Table 3-16 show that the improvements obtained with the term deletion process are fully significant.

Table 3-17 includes a comparison of various composite term weighting systems such as those based on inverse document frequency weighting ($\text{FREQ}_{ik}/\text{DOCFREQ}_k$) [equivalent to expression (2)], on the signal value $\text{FREQ}_{ik} \cdot \text{SIGNAL}_k$ [expression (6)] and on the discrimination value $\text{FREQ}_{ik} \cdot \text{DISCVALUE}_k$ [expression (9)]. It may be seen that the composite weights obtained with the inverse document frequency and the discrimination value weights are approximately comparable in efficiency. The results produced by the signal value are substantially less attractive because of the emphasis on low-frequency terms inherent in that weighting system. The two right-hand columns of Table 3-17a demonstrate that the composite weights combined with the elimination of poor terms produces superior precision values at low and medium recall levels. The significance test results of Table 3-17b indicate fully significant performance improvements with an average performance difference ranging from 8 to 15 percent.

A wide variety of phrase-generation methods are potentially useful, as explained earlier. To generate the output of Table 3-18 a simple phrase-generation method was used consisting of phrases obtained from pairs (P) and triples (T) of co-occurring nondiscriminators [104]. Specifically, given three nondiscriminators T_i, T_j, and T_k occurring in a given document abstract, one triple T_{ijk} can be formed as well as three pairs T_{ij}, T_{ik}, and T_{jk}. For the *Time* collection under study an average of 11 phrases were generated by this simple process for each document. Single terms (S), pairs (P), and triples (T) can all be used together (denoted SPT); alternatively, pairs and triples alone can be added to the vec-

Recall	Standard term frequency FREQ_{ik} weights	FREQ_{ik} weights with IDF $\text{FREQ}_{ik}/\text{DOCFREQ}_k$	FREQ_{ik} weights with DISCVALUE_k $\text{FREQ}_{ik} \cdot \text{DISCVALUE}_k$	FREQ_{ik} weights with SIGNAL_k $\text{FREQ}_{ik} \cdot \text{SIGNAL}_k$	$\text{FREQ}_{ik}/\text{DOCFREQ}_k$ with IDF CUT	$\text{FREQ}_{ik} \cdot \text{DISCVALUE}_k$ with DISC CUT
0.1	0.7496	0.8536	0.8406	0.7212	0.8975	0.8028
0.2	0.7071	0.7901	0.7881	0.7006	0.8315	0.7480
0.3	0.6710	0.7568	0.7197	0.6471	0.7800	0.7286
0.4	0.6452	0.7305	0.6901	0.6229	0.7574	0.6938
0.5	0.6351	0.6783	0.6704	0.6105	0.7372	0.6737
0.6	0.5866	0.6243	0.6176	0.5587	0.6529	0.6349
0.7	0.5413	0.5823	0.5727	0.5263	0.5912	0.5847
0.8	0.5004	0.5643	0.5169	0.4612	0.5481	0.5475
0.9	0.3865	0.4426	0.4208	0.3830	0.4318	0.4259
1.0	0.3721	0.4170	0.4053	0.3593	0.4118	0.4085

(a)

	t-test	Wilcoxon		t-test	Wilcoxon
A. FREQ_{ik} weight with IDF $\text{FREQ}_{ik}/\text{DOCFREQ}_k$	0.0000	0.0000	A. $\text{FREQ}_{ik}/\text{DOCFREQ}_k$ with IDF CUT	0.0000	0.0000
B. Standard FREQ_{ik}	A better than B by 11%		B. Standard FREQ_{ik}	A better than B by 15%	
A. FREQ_{ik} weight with DISCVALUE_k	0.0000	0.0000	A. $\text{FREQ}_{ik} \cdot \text{DISCVALUE}_k$ with DISC CUT	0.0084	0.0077
B. Standard FREQ_{ik}	A better than B by 8%		B. Standard FREQ_{ik}	A better than B by 8%	

(b)

Table 3-17 Composite weighting functions (425 documents, 24 queries). (a) Comparison of composite weighting systems. (b) Statistical significance results for output of Table 3-17a (testing for A better than B).

Recall	FREQ_{ik} control run	Best frequency weighting $\mathrm{FREQ}_{ik}/\mathrm{DOCFREQ}_k$	Best phrase process PT+SPT	Thesaurus classes	Thesaurus + PT+SPT
0.1	0.7496	0.8536	0.8860	0.7392	0.8761
0.2	0.7071	0.7901	0.7984	0.7166	0.7972
0.3	0.6710	0.7568	0.7761	0.6935	0.7778
0.4	0.6452	0.7305	0.7461	0.6627	0.7465
0.5	0.6351	0.6783	0.7020	0.6541	0.7027
0.6	0.5866	0.6243	0.6563	0.6070	0.6524
0.7	0.5413	0.5823	0.6010	0.5598	0.6010
0.8	0.5004	0.6543	0.5483	0.5111	0.5523
0.9	0.3865	0.4426	0.4231	0.4091	0.4260
1.0	0.3721	0.4170	0.4118	0.3950	0.4149

(a)

	t-test	Wilcoxon
A. Thesaurus + PT + SPT phrases	0.6874	0.6833
B. $\mathrm{FREQ}_{ik}/\mathrm{DOCFREQ}_k$ weights		
A. Thesaurus + PT + SPT phrases	0.4524	0.9657
B. PT + SPT phrases		
A. Thesaurus	0.0000	0.0003
B. Standard term frequency FREQ_{ik}		

(b)

Table 3-18 Thesaurus and phrase evaluation. (a) Thesaurus and phrase performance. (b) Statistical significance results for output of Table 4-17a (testing for A better than B).

tors, the corresponding single terms being deleted (PT). When high-frequency nondiscriminators are used for phrase generation, the PT method appears to offer a reasonably high performance standard.

A manually constructed thesaurus designed to group low-frequency terms (terms with a document frequency smaller than 20 out of 425 documents) was also available for experimental purposes with the *Time* collection; the relevant thesaurus class identifiers can be added in each case to the standard document and query vectors.

The performance data are included in Table 3-18 together with the corresponding statistical significance output. It may be seen that for the *Time* collection, the grouping of low-frequency terms into thesaurus classes affords recall improvements over the standard FREQ_{ik} weighting system for all but the lowest recall levels. Moreover, the thesaurus advantage proves statistically significant. The phrase-generation process, however, proved more effective especially at the high precision–low recall end of the performance spectrum. In the middle recall range, the best performance of any of the retrieval runs displayed in Tables 3-15 to 3-18 is obtained by using a combination of the phrase-generation process applied to high-frequency terms with the thesaurus class grouping of low-frequency terms.

The evaluation results presented in Tables 3-15 to 3-18 are indicative of the performance of various sophisticated statistically based automatic indexing methods. Substantial additional work remains to be done in order to determine the optimum indexing system applicable to a particular retrieval environment under given conditions. When no special information is available about a particular collection, the following process will, however, provide a high-quality indexing product:

1 Starting with document abstracts or excerpts, remove common high-frequency words and generate word stems by removing suffixes from the remaining words.

2 Compute the discrimination values of the terms, generate phrases for the high-frequency nondiscriminators with negative discrimination values, and assemble low-frequency nondiscriminators with near-zero discrimination values into thesaurus classes.

3 Compute a weighting factor for each remaining single term, phrase, and thesaurus class, using, for example, the inverse document frequency function [expression (2)].

4 Assign to each document the weighted term vector consisting of single terms, phrases, and thesaurus classes.

In interactive retrieval systems where user information is available during the search process, improved index term assignments may be obtainable as a result of the user-system interaction. The possibility is explored further in Chapter 6.

REFERENCES

[1] E.M. Keen, On the Generation and Searching of Entries in Printed Subject Indexes, Journal of Documentation, Vol. 33, No. 1, March 1977, pp. 15–45.

[2] B.C. Vickery, Techniques of Information Retrieval, Archon Books, Hamden, Connecticut, 1970.

[3] E.M. Keen, On the Performance of Nine Printed Subject Index Entry Types, Research Report, College of Librarianship, Aberystwyth, Wales, September 1978.

[4] F.W. Lancaster, Information Retrieval Systems, Characteristics, Testing and Evaluation, 2nd Edition, Wiley-Interscience, New York, 1979.

[5] W.S. Cooper, Is Interindexer Consistency a Hobgoblin?, American Documentation, Vol. 20, No. 3, July 1969, pp. 268–278.

[6] P. Zunde and M.E. Dexter, Indexing Consistency and Quality, American Documentation, Vol. 20, No. 3, July 1969, pp. 259–267.

[7] M.E. Stevens, Automatic Indexing—A State of the Art Report, National Bureau of Standards, NBS Monograph 91, National Bureau of Standards, Washington, D.C., March 1965.

[8] H.P. Luhn, The Automatic Creation of Literature Abstracts, IBM Journal of Research and Development, Vol. 2, No. 2, April 1958, pp. 159–165.

[9] G.K. Zipf, Human Behavior and the Principle of Least Effort, Addison Wesley Publishing, Reading, Massachusetts, 1949.

[10] H. Kucera and W.N. Francis, Computational Analysis of Present-Day American English, Brown University Press, Providence, Rhode Island, 1967.

[11] H.P. Luhn, A Statistical Approach to Mechanized Encoding and Searching of Literary Information, IBM Journal of Research and Development, Vol. 1, No. 4, October 1957, pp. 309–317.

[12] H.P. Edmundson and R.E. Wyllys, Automatic Abstracting and Indexing—Survey and Recommendations, Communications of the ACM, Vol. 4, No. 5, May 1961, pp. 226–234.

[13] F.J. Damerau, An Experiment in Automatic Indexing, American Documentation, Vol. 16, No. 4, October 1965, pp. 283–289.

[14] K. Sparck Jones, A Statistical Interpretation of Term Specificity and Its Application in Retrieval, Journal of Documentation, Vol. 28, No. 1, March 1972, pp. 11–20.

[15] C.E. Shannon, Prediction and Entropy of Printed English, Bell System Technical Journal, Vol. 30, No. 1, January 1951, pp. 50–65.

[11] S.F. Dennis, Law, Language, Words, Entropy, and Automatic Indexing, unpublished manuscript.

[11] S.F. Dennis, The Design and Testing of a Fully Automatic Indexing-Searching System for Documents Consisting of Expository Text, in Information Retrieval: A Critical Review, G. Schecter, editor, Thompson Book Co., Washington, D.C., 1967, pp. 67–94.

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

[19] 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.

[20] G. Salton and M.E. Lesk, Recent Studies in Automatic Text Analysis and Information Retrieval, Journal of the ACM, Vol. 20, No. 2, April 1973, pp. 258–278.

[21] G. Salton, C.S. Yang, and C.T. Yu, A Theory of Term Importance in Automatic Text Analysis, Journal of the American Society for Information Science, Vol. 26, No. 1, January–February 1975, pp. 33–44.

[22] 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.

[23] R.A. May, editor, Automated Law Research, American Bar Association, Chicago, Illinois, 1973.

[24] G. Salton and M.E. Lesk, Computer Evaluation of Indexing and Text Processing, Journal of the ACM, Vol. 25, No. 1, January 1968, pp. 8–36.

[25] C.J. van Rijsbergen, Information Retrieval, 2nd Edition, Butterworths, London, 1979.

[26] G. Salton, Automatic Information Organization and Retrieval, McGraw-Hill Book Company, New York, 1968.

[27] J.B. Lovins, Development of a Stemming Algorithm, Mechanical Translation and Computational Linguistics, Vol. 11, No. 1–2, March and June 1968, pp. 11–31.

[28] G. Salton and M.E. Lesk, Information Analysis and Dictionary Construction, in

the SMART Retrieval System—Experiments in Automatic Document Processing, G. Salton, editor, Chapter 6, Prentice-Hall, Inc., Englewood Cliffs, New Jersey, 1971.

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

[30] G. Salton, Generation and Search of Clustered Files, ACM Transactions on Data Base Systems, Vol. 3, No. 4, December 1978, pp. 321–346.

[31] 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, New Jersey, 1971.

[32] G. Salton, Dynamic Information and Library Processing, Prentice-Hall, Inc., Englewood Cliffs, New Jersey, 1975.

[33] V.E. Giuliano and P.E. Jones, Linear Associative Information Retrieval, in Vistas in Information Handling, P. Howerton, editor, Spartan Books, Inc., Washington, D.C., 1963.

[34] M.E. Lesk, Word-Word Associations in Document Retrieval Systems, American Documentation, Vol. 20, No. 1, January 1969, pp. 27–38.

[35] L.B. Doyle, Information Retrieval and Processing, Melville Publishing Company, Los Angeles, California, 1975.

[36] G. Jones and P. Wise, Updating Thesaurus Classifications in Response to Data Base Changes, Scientific Report No. ISR-21, Section XIII, Department of Computer Science, Cornell University, Ithaca, New York, December 1972.

[37] M.M. Kessler, Comparison of the Results of Bibliographic Coupling and Analytic Subject Indexing, American Documentation, Vol. 16, No. 3, 1965, pp. 223–233.

[38] G. Salton, Automatic Indexing Using Bibliographic Citations, Journal of Documentation, Vol. 27, No. 2, June 1971, pp. 98–110.

[39] H.P. Edmundson, Problems in Automatic Abstracting, Communications of the ACM, Vol. 7, No. 4, April 1964, pp. 259–263.

[40] H.P. Edmundson, New Methods in Automatic Extracting, Journal of the ACM, Vol. 26, No. 2, April 1969, pp. 264–285.

[41] J.E. Rush, R. Salvador, and A. Zamora, Automatic Abstracting and Indexing II, Production of Indicative Abstracts by Application of Contextual Inference and Syntactic Coherence Criteria, Journal of the ASIS, Vol. 22, No. 4, July–August 1971, pp. 260–274.

[42] C.A. Montgomery, Linguistics and Information Science, Journal of the American Society for Information Science, Vol. 23, No. 3, May–June 1972, pp. 195–219.

[43] K. Sparck Jones and M. Kay, Linguistics and Information Science, Academic Press, New York, 1973.

[44] F. Damerau, Automated Language Processing, in Annual Review of Information Science and Technology, M.E. Williams, editor, American Society for Information Science, Washington, D.C., Vol. 11, 1976, pp. 107–161.

[45] S. Kuno and A. G. Oettinger, Multiple-Path Syntactic Analyzer, in Information Processing 62, North Holland Publishing Company, Amsterdam, 1963, pp. 128–133.

[46] P. Baxendale, An Empirical Model for Machine Indexing, in Machine Indexing—Progress and Problems, Third Institute on Information Storage and Retrieval, American University, February 1961, pp. 207–218.

[47] W.D. Climenson, N.H. Hardwick, and S.N. Jacobson, Automatic Syntax Anal-

ysis in Machine Indexing and Abstracting, American Documentation, Vol. 12, No. 3, July 1961, pp. 178–183.

[48] F.J. Damerau, Automatic Parsing for Content Analysis, Communications of the ACM, Vol. 13, No. 6, June 1970, pp. 356–360.

[49] D.J. Hillman and A.J. Kasarda, The LEADER Retrieval System, AFIPS Proceedings, AFIPS Press, Montvale, New Jersey, Vol. 34, 1969, pp. 447–455.

[50] G. Salton, Automatic Phrase Matching, in Readings in Automatic Language Processing, D. Hays, editor, American Elsevier Publishing Company, New York, 1966, pp. 169–188.

[51] J. Friedman, A Computer System for Transformational Grammar, Communications of the ACM, Vol. 12, No. 6, June 1969, pp. 341–348.

[52] N. Sager and R. Grishman, The Restriction Language for Computer Grammars of Natural Language, Communications of the ACM, Vol. 18, No. 7, July 1975, pp. 390–400.

[53] W.A. Woods, Transition Network Grammars for Natural Language Analysis, Communications of the ACM, Vol. 13, No. 10, October 1970, pp. 591–606.

[54] G.G. Hendrix, Human Engineering for Applied Natural Language Processing, Fifth International Joint Conference on Artificial Intelligence, M.I.T., Cambridge, Massachusetts, 1977, pp. 183–191.

[55] L.R. Harris, User-Oriented Data Base Query with the ROBOT Natural Language Query System, International Journal of Man-Machine Studies, Vol. 9, 1977, pp. 697–713.

[56] J. Mylopoulos, A. Borgida, P. Cohen, N. Roussopoulos, J. Tsotsos and H. Wong, Torus—A Natural Language Understanding System for Data Management, Fourth International Joint Conference on Artificial Intelligence, Tbilisi, USSR, September 1975, pp. 414–421.

[57] W.J. Plath, REQUEST: A Natural Language Question-Answering System, IBM Journal of Research and Development, Vol. 20, No. 4, July 1976, pp. 326–335.

[58] D.L. Waltz, An English Language Question Answering System for a Large Relational Database, Communications of the ACM, Vol. 21, No. 7, July 1978, pp. 526–539.

[59] G.S. Dunham, M.G. Pacak, and A.W. Pratt, Automatic Indexing of Pathology Data, Journal of the American Society for Information Science, Vol. 29, No. 2, March 1978, pp. 81–90.

[60] R. Grishman and L. Hirschman, Question Answering from Natural Language Medical Data Bases, Artificial Intelligence, Vol. 11, No. 1/2, 1978, pp. 25–43.

[61] H.J. Schek, The Reference String Indexing Method, Research Report, IBM Scientific Center, Heidelberg, Germany, 1978.

[62] E.J. Schuegraf and H.S. Heaps, Query Processing in a Retrospective Document Retrieval System That Uses Word Fragments as Language Elements, Information Processing and Management, Vol. 12, No. 4, 1976, pp. 283–292.

[63] E.J. Schuegraf and H.S. Heaps, Selection of Equifrequent Word Fragments for Information Retrieval, Information Storage and Retrieval, Vol. 9, No. 12, December 1973, pp. 697–711.

[64] M.F. Lynch, Variety Generation—A Reinterpretation of Shannon's Mathematical Theory of Communication and Its Implications in Information Science, Journal of the American Society for Information Science, Vol. 28, No. 1, January 1977, pp. 19–25.

[65] T. Radhakrishnan, Selection of Prefix and Postfix Word Fragments for Data Com-

pression, Information Processing and Management, Vol. 14, No. 2, 1978, pp. 97–106.

[66] I.J. Barton, S.E. Creasey, M.F. Lynch, and M.J. Snell, An Information-Theoretic Approach to Text Searching in Direct Access Systems, Communications of the ACM, Vol. 17, No. 6, June 1974, pp. 345–350.

[67] H.S. Heaps, Information Retrieval—Computational and Theoretical Aspects, Academic Press, New York, 1978.

[68] D. Gottlieb, S.A. Hagerth, P.G.H. Lehot, and H.S. Rabinowitz, A Classification of Compression Methods and Their Usefulness for a Large Data Processing Center, AFIPS Conference Proceedings, Vol. 44, 1975, pp. 453–458.

[69] M. Kochen, Principles of Information Retrieval, Melville Publishing Company, Los Angeles, California, 1974.

[70] W. Goffman, A Searching Procedure for Information Retrieval, Information Storage and Retrieval, Vol. 2, No. 2, July 1964, pp. 73–78.

[71] A. Bookstein and D.R. Swanson, A Decision Theoretic Foundation for Indexing, Journal of the ASIS, Vol. 26, No. 1, January–February 1975, pp. 45–50.

[72] C.J. van Rijsbergen, A. Theoretical Basis for the Use of Co-occurrence Data in Information Retrieval, Journal of the Documentation, Vol. 33, No. 2, June 1977, pp. 106–119.

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

[74] F.W. Lancaster and E.G. Fayen, Information Retrieval On-Line, John Wiley and Sons, New York, 1973.

[75] M.E. Lesk and G. Salton, Interactive Search and Retrieval Methods Using Automatic Information Displays, in the SMART Retrieval System, Experiments in Automatic Document Processing, G. Salton, editor, Prentice-Hall, Inc., Englewood Cliffs, New Jersey, 1971, Chapter 25.

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

[77] S.E. Robertson and K. Sparck Jones, Relevance Weighting of Search Terms, Journal of the ASIS, Vol. 27, No. 3, May–June 1976, pp. 129–146.

[78] 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.

[79] G. Salton and R.K. Waldstein, Term Relevance Weights in On-Line Information Retrieval, Information Processing and Management, Vol. 14, No. 1, 1978, pp. 29–35.

[80] M.E. Maron, Automatic Indexing—An Experimental Inquiry, Journal of the ACM, Vol. 8, No. 3, July 1961, pp. 404–417.

[81] M.E. Maron and J.L. Kuhns, On Relevance, Probabilistic Indexing and Information Retrieval, Journal of the ACM, Vol. 7, No. 3, July 1960, pp. 216–244.

[82] W.B. Croft, A Study of the Effectiveness and Implementation of a Model of Cluster Searching, Research Report, University of Cambridge, Computer Laboratory, 1978.

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

[84] F. Mosteller and E.L. Wallace, Inference in an Authorship Problem, Journal of the American Statistical Association, Vol. 58, No. 302, June 1963, pp. 275–309.

[85] S.P. Harter, A Probabilistic Approach to Keyword Indexing, Part 1. On the Distribution of Specialty Words in a Technical Literature, Journal of the ASIS, Vol. 26, No. 4, July–August 1975, pp. 197–206.

[86] A. Bookstein and D. Kraft, Operations Research Applied to Document Indexing and Retrieval Decisions, Journal of the ACM, Vol. 24, No. 3, July 1977, pp. 418–427.

[87] G. Salton, Automatic Text Analysis, Science, Vol. 168, No. 3929, April 1970, pp. 335–343.

[88] C. Montgomery and D.R. Swanson, Machine-like Indexing by People, American Documentation, Vol. 13, No. 4, October 1962, pp. 359–366.

[89] M.J. Ruhl, Chemical Documents and Their Titles: Human Concept Indexing vs. KWIC Machine Indexing, American Documentation, Vol. 15, No. 2, April 1964, pp. 136–141.

[90] D.H. Kraft, A Comparison of Keyword in Context (KWIC) Indexing of Titles with a Subject Heading Classification System, American Documentation, Vol. 15, No. 1, January 1964, pp. 48–52.

[91] M.E. Stevens and G.H. Urban, Training a Computer to Assign Descriptors to Documents: Experiments in Automatic Indexing, Proceedings of the Spring Joint Computer Conference, Spartan Books, 1964, pp. 563–575.

[92] T.N. Shaw and H. Rothman, An Experiment in Indexing by Word Choosing, Journal of Documentation, Vol. 24, No. 3, September 1968, pp. 159–172.

[93] D.R. Swanson, Searching Natural Language Text by Computer, Science, Vol. 132, No. 3434, October 21, 1960, pp. 1099–1104.

[94] D.R. Swanson, Interrogating a Computer in Natural Language, in Information Processing 62 (Proceedings of IFIP Congress 62), C. Popplewell, editor, North Holland Publishing Co., Amsterdam, 1963, pp. 288–293.

[95] T. Saracevic, An Inquiry into Testing of Information Retrieval Systems, Comparative Systems Laboratory Technical Reports No. CSL: TR-FINAL 1 to 3, Center for Documentation and Communication Research, Case Western Reserve University, Cleveland, Ohio, 1968.

[96] C.W. Cleverdon and E.M. Keen, Factors Determining the Performance of Indexing Systems, Vol. 1: Design, Vol. 2: Test Results, Aslib Cranfield Research Project, Cranfield, England, 1966.

[97] C.W. Cleverdon, The Cranfield Tests on Index Language Devices, Aslib Proceedings, Vol. 19, No. 6, June 1967, pp. 173–194.

[98] G. Salton, editor, the SMART Retrieval System—Experiments in Automatic Document Processing, Prentice-Hall, Inc., Englewood Cliffs, New Jersey, 1971.

[99] G. Salton, editor, Scientific Reports No. ISR-11 to ISR-22, Department of Computer Science, Cornell University, Ithaca, New York, 1966–1974.

[100] F.W. Lancaster, Evaluation of the MEDLARS Demand Search Service, National Library of Medicine, Bethesda, Maryland, January 1968.

[101] G. Salton, A Comparison between Manual and Automatic Indexing Methods, American Documentation, Vol. 20, No. 1, January 1969, pp. 61–71.

[102] G. Salton, A New Comparison between Conventional Indexing (MEDLARS) and Automatic Text Processing (SMART), Journal of the American Society for Information Science, Vol. 23, No. 2, March–April 1972, pp. 75–84.

[103] R.W. Crawford, Negative Dictionary Construction, Scientific Report No. ISR-22, Department of Computer Science, Cornell University, Ithaca, New York, November 1974.

[104] G. Salton and A. Wong, On the Role of Words and Phrases in Automatic Text Analysis, Computers and the Humanities, Vol. 10, No. 2, March–April 1976, pp. 69–87.

BIBLIOGRAPHIC REMARKS

For a review of conventional tools for document indexing see:

B.C. Vickery, Techniques of Information Retrieval, Archon Books, Hamden, Connecticut, 1970.

D. Soergel, Indexing Languages and Thesauri: Construction and Maintenance, Melville Publishing Company, Los Angeles, California, 1974.

Additional basic information on automatic indexing methods may be obtained from the following references:

M.E. Stevens, Automatic Indexing: A State of the Art Report, Monograph 91, National Bureau of Standards, Washington, D.C., March 1965.

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

L.B. Doyle, Information Retrieval and Processing, Melville Publishing Company, Los Angeles, California, 1975.

EXERCISES

3-1 Explain the significance of each of the following indexing procedures, and determine the effect of each method as a means of enhancing recall and/or precision, respectively:

- **a** Word stemming process
- **b** Use of synonym dictionary
- **c** Use of word location information
- **d** Use of term weights

3-2 What is the significance of term frequency in the theory of indexing? What is the significance of Zipf's law as a basis for deriving an automatic indexing method? Is the discrimination value of a term related to the occurrence frequency of a term in a document collection? If so, what is the relationship?

3-3 The term relevance factor TERMREL differs from all other term weighting systems in the sense that relevance information must be available for some documents with respect to a given query. Explain the role of relevance information for the computation of the TERMREL weight and give two methods for estimating the required relevance information.

3-4 Consider the following sample document collection:

$$
\begin{aligned}
D_1 &= (1,0,1,0,0,0)\\
D_2 &= (3,1,2,1,0,1)\\
D_3 &= (1,2,3,0,1,0)\\
D_4 &= (0,1,0,2,1,2)\\
D_5 &= (1,0,1,4,2,1)\\
D_6 &= (1,1,0,2,3,2)
\end{aligned}
$$

Generate a term-term similarity matrix similar to that shown in Table 4-10 using one of the term similarity coefficients given in expressions (11) or (12). Choose a threshold value for the term similarity and "expand" the sample documents by adding associated terms to the original term vectors. In what respects do the new expanded vectors differ from the originals?

3-5 Consider the document collection of Exercise 3-4 together with the following query pair

$$Q_1 = (2,0,2,0,0,0)$$
$$Q_2 = (0,0,0,2,0,2)$$

a Exhibit the normal query-document similarity coefficients for all query-document pairs using a vector similarity function such as expressions (11) and (12) in the form SIMILAR (D_i, Q_j). Display the documents in decreasing order of query similarity.

b Compute the discrimination value for each term, and construct updated document vectors using discrimination value weighting.

c Repeat part a using the discrimination value weighting.

d Assuming that D_1, D_2, D_3 are relevant to Q_1 and D_4, D_5, D_6 are relevant to Q_2, compute the term relevance value (TERMREL) for each term and construct updated vectors using term relevance value weighting.

e Repeat part a using the relevance weighting.

f Repeat part a for the "expanded" collection derived in Exercise 3-4.

g Compare the results of parts a, c, e, and f.