Automatic Text Analysis

Automatic document indexing and classification methods are examined and their effectiveness is assessed.

G. Salton

Over the years, linguists and philosophers of language have deplored the fact that no adequate theory exists to account for many of the important phenomena connected with the natural language. The lack of such a theory has in particular made it impossible to construct linguistic models which would completely and accurately represent the structure of the natural language, and this in turn has led to predictions that computers, which must rely on some model specifying the processing rules, would have only a relatively minor role in the analysis of written texts (*1*).

While remarks concerning the lack of appropriate linguistic theories and models are entirely justified, the many experiments conducted over the last few years in the general area of automatic text processing, including automatic document indexing and classification and automatic text analysis, provide evidence that computer-based text processing is practicable and useful for many kinds of applications. In fact, the indication is that some of the automatic content analysis and text processing methods not only are relatively easy to implement but can be used in an automatic information storage and retrieval environment to produce a retrieval effectiveness at least equal to that obtained by the conventional, mostly manual procedures used in the past.

In this article the principal experiments in automatic text analysis are briefly reviewed, and an indication is given of developments to be expected in the future.

General Methodology

The first serious work in automatic text analysis dates back to the middle and late 1950's, when Luhn argued that the vocabulary contained in individual document texts would necessarily have to constitute the basis for a useful content analysis and classification (*2, 3*). Several possible indexing methods were proposed by Luhn, including, for example, the following (*2*, p. 315):

> . . . 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 paragraph 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.

Luhn further suggested that the inquirer's "document" (that is, the search request) be encoded in exactly the same manner as the documents of the collection, so that queries and documents could appropriately be matched.

These early ideas were not universally appreciated, partly because they could not be applied uniformly to all query and document texts—many counterexamples were produced to show that a given methodology would not operate under certain circumstances—and partly because the automatic procedures were never adequately tested. Nevertheless, a good deal of work has been done to refine and expand the original ideas, and several operational automatic content analysis systems are now in existence. The following types of operations are often used.

1) Expressions are first chosen from the document or query texts; often this implies the identification or generation of words, word stems, noun phrases, prepositional phrases, or other content units, with certain specified properties.

2) A weight may be assigned each expression on the basis of the frequency of occurrence of the given expression, or the position of the expression in the document, or the type of entity.

3) Expressions originally assigned to documents may be replaced by new expressions, or new expressions may be added to those originally available, thereby "expanding" the set of content identifiers; such an expansion may be based on information contained in a stored dictionary, or, alternatively, it may be based on statistical co-occurrence characteristics between the terms in a document collection, or on syntactical relations between words.

4) Additional relational indicators between expressions may be supplied to express syntactical, or functional, or logical relationships between the entities available for content identification.

The result of such an automatic indexing process is then similar to that outlined by Luhn in the sense that each document or search request is identified by a set of terms. However, these terms may consist of complete phrases which do not necessarily originate in the document to which they are assigned; moreover, each term may carry a weight reflecting its presumed importance for purposes of content analysis.

It is impossible in a brief article such as this to discuss the many strategies that have been proposed for automatic indexing (*4*). Instead, the experimental evidence derived from many of the recent studies in automatic indexing and text analysis is examined, and conclusions are drawn concerning the effectiveness of the various techniques.

Indexing Experiments

Most of the early experiments in automatic indexing did not include any kind of retrieval test but consisted principally of a comparison between automatically derived index terms and preestablished manually assigned subject categories. Typically, a manually indexed document collection would be taken, and an attempt would be made to duplicate by automatic means as many of the preassigned terms as possible. Three types of studies may be distinguished, depending on the testing device actually used: (i) title word studies, (ii) studies involving comparisons of automatically generated and manually assigned terms, and (iii) studies based on automatic assignment to known subject classes.

The title word studies use as a criterion the similarity between entries derived from document titles and from manually assigned subject headings. Montgomery and Swanson (*5*) used an issue of *Index Medicus* containing title

The author is professor of computer science, Cornell University, Ithaca, New York.

citations in biomedicine cross-filed under the various manually derived subject headings and concluded that, for 86 percent of almost 5000 titles, a correlation existed between the subject heading assigned in *Index Medicus* and the document title ("Correlation" was defined as an actual match between word stems, or a match between rather loosely defined synonymous terms). In a somewhat related study based on use of the chemical literature, Ruhl (*6*) found that 57 percent of the titles examined contained all the important concepts (or their equivalents) listed for these documents in the subject index of *Chemical Abstracts*, while only 12 percent of the titles missed three or more important subject headings. Similar results were found by Kraft (*7*) for the legal literature: only about 10 percent of the document titles examined did not contain any key words useful for indexing purposes, while 64 percent of the title entries contained one or more of the subject heading words included in the *Index to Legal Periodicals*, and an additional 25 percent of the titles contained "logical equivalents" of the subject headings.

While results of this type are not directly usable, particularly in the absence of tests in a retrieval environment, the evidence nevertheless suggests that simple automatic word-extracting methods are not necessarily worthless. Furthermore, the counter-evidence cited by O'Connor (*8*), who finds a correlation between assigned subject headings and title words ranging from a low of only 13 percent to a high of only 68 percent, was produced with a very strict definition of synonymy (that is, the terms were required to be strictly synonymous in order to be considered equivalent) which is not necessarily desirable either for indexing or for retrieval (*8*).

The next set of experiments consists of a comparison between automatically generated and manually assigned sets of index terms. Such term set comparisons are often performed by matching a set of automatically generated index terms with the available manually assigned terms. An evaluation coefficient such as q may be used to measure the amount of overlap between vocabularies, where

$$q = \frac{c}{a + m - c}$$

[Here c represents the number of common term assignments, a is the number of automatically derived terms, and m is the number of manually assigned terms (*9*).]

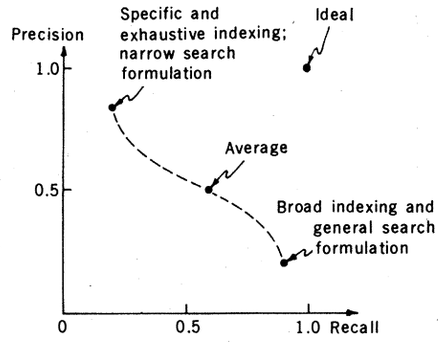

Fig. 1. Typical recall-precision graph reflecting performance characteristics of retrieval systems.

Various tests of this general type have been performed (*10, 11*), and the consensus is that about 60 percent agreement between manually and automatically produced terms is obtainable. In one test involving the automatic assignment of phrases to documents, as many as 86 percent of the automatically assigned phrases were found by human judges to be acceptable subject heads, the "overassignment" (that is, the assignment of extraneous phrases) being of the order of 14 percent, and the "underassignment" (that is, proper content indicators not recognized by the machine) being of the order of 11 percent (*12*). A related approach, consisting of a comparison between automatically derived document affinities based on similarities in the bibliographic citations attached to the documents with document affinities based on overlapping sets of manually assigned subject headings, also indicates a considerable amount of agreement between the automatic and manual procedures (*13*).

The last set of experiments (apart from tests in a retrieval environment) involves automatic classification of documents into subject categories (rather than assignment of index terms to documents) (*14*). This is often accomplished as follows. A *test* collection is used manually to classify documents into subject categories and to compute similarity parameters between a given subject category and the vocabularies of documents contained in that category. These parameters are then used automatically to classify a *control* collection consisting of new, incoming documents (*15, 16*). It is found that, for the original test documents, an automatic assignment to subject categories is about 80 to 90 percent effective (that is, the correct category is chosen in about 80 to 90 percent of the cases). For control items—documents not used in deriving the test parameters—the effectiveness of the automatic classification based on document vocabularies drops down to about 50 percent.

O'Connor (*16*) remarks that the percentage of correctly classified documents increases when more refined classification parameters are used (from 76 percent when key words alone are used to 92 percent when certain relationships between key words are also utilized); at the same time, the number of incorrectly classified items which are wrongly included in a category also increases from 13 to 18 percent. This tradeoff between the number of correct and incorrect responses—as the first goes up, the second goes up also—is characteristic of retrieval system performance.

Retrieval Experiments

The indexing experiments described above were not performed within a normal retrieval situation, and involved reliance on criteria supplied by human experts for purposes of evaluation. In a retrieval environment, on the other hand, it is possible to use as a criterion of system effectiveness the ability of the system to satisfy the user's need for information by retrieving wanted material and rejecting unwanted items. Two measures have been widely used for this purpose, known as "recall" and "precision," and representing, respectively, the proportion of relevant material actually retrieved and the proportion of retrieved material actually relevant. Ideally, all relevant items should be retrieved and all nonrelevant items should be rejected; such a situation is reflected in perfect recall and precision values, equal to 1.

It should be noted that both the recall and the precision figures achievable by a given system are adjustable, in the sense that a relaxation of search criteria (a broader search formulation) often leads to high recall, while a tightening of search criteria (a narrower search formulation) leads to high precision. Unhappily, experience has shown that, on the average, recall and precision tend to vary inversely: the retrieval of a greater number of relevant items normally also leads to the retrieval of a greater number of irrelevant ones. When recall and precision are plotted against each other, a monotonically decreasing curve of

the type shown in Fig. 1 reflects the average performance characteristic of a retrieval system.

In practice, a compromise is usually made, and a performance level is chosen such that much of the relevant material is retrieved while the number of nonrelevant items also retrieved is kept within tolerable limits. Thus, in what is probably the most exhaustive evaluation of an operating retrieval system involving the use of manually indexed documents (in this case the Medlars system at the National Library of Medicine), Lancaster (*17*) reports an average recall of 0.577 and an average precision of 0.504 (*18*). Additional evaluation results are to be found in the extensive literature dealing with the evaluation of operating, manually based retrieval systems (*19*).

The first comparison of conventional retrieval (retrieval of manually indexed documents) with automatic text processing systems appears to be the one made by Swanson in the late 1950's, using 100 documents and 50 queries (*20*). Three indexing and analysis systems were used: (i) conventional retrieval based on a subject heading index; (ii) retrieval based on specifications provided by words and phrases automatically extracted from the document texts; and (iii) retrieval based on use of a thesaurus in addition to the words obtained from the documents. A measure similar to the "recall-precision" measure was used to evaluate system performance; it varied directly with the relevance weight of the retrieved items and included in addition a penalty factor for retrieval of irrelevant material.

The test results indicated that the average retrieval performance of a system based on automatic text analysis was superior to the performance of the standard system based on manual indexing. Since Swanson provides the first of a long series of results all tending to prove the same point, it is worth quoting from his report (*20*):

> The first conspicuous implication of the result is that the proportion of relevant information retrieved under any circumstances is rather low.
>
> The second implication of the data is the apparent superiority of machine-retrieval techniques over conventional retrieval within the famework of our model. Conventional retrieval was carried out under the favorable conditions of a highly detailed and specific subject-heading list, tailored to a sample library. . . .
>
> It is expected that the relative superiority of machine text searching to conventional retrieval will become greater with subsequent experimentation as retrieval aids for text searching are improved, whereas no clear procedure is in evidence which will guarantee improvement of the conventional system. . . . Thus even though machines may never enjoy more than a partial success in library indexing, a small suspicion might justifiably be entertained that people are even less promising.

In view of the test results produced by far more extensive experimentation reported below, these prophesies appear to have been remarkably accurate.

Swanson's original results were confirmed in an extension of the test in which, for the first time, natural-language queries were used (instead of manually constructed query formulations) (*21*). 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 then possible to compute recall and precision values following the retrieval of each document (or each *n*th document); this produces a sequence of recall-precision pairs which can be plotted as a curve similar to that of Fig. 1. Swanson concludes his study by stating (*21*):

> . . . though these results [that automatic text processing using an automatic thesaurus is more accurate than the human process of assigning appropriate subject index terms to documents and queries] may violate one's sense of intuition, there is no good theoretical reason to believe that they ought to have come out differently.

Various later studies also included elements of automatic text analysis, including full text search (*22*), the use of phrase dictionaries and syntactic analysis procedures (*23, 24*), statistical term associations (*25, 26*), and automatically constructed term groupings (thesauri) (*27*). In each case the intent is to show that one or another of the proposed automatic language analysis methods operates more successfully than either a manual indexing process or an automatic process based on a less sophisticated approach. In general, the case is made that the use of manually constructed thesauri or of automatic term associations or term groups is useful in a retrieval environment. [In the one case where an automatic phrase-matching procedure appeared not to produce reasonable results, the test conditions were peculiar, since the texts processed in the experiment were not the same as those used to determine the relevance of a given document to a query (*24*); furthermore, retrieval appears to have been based on the presence, or absence, of a single matching phrase or sentence fragment, so the test results are difficult to interpret effectively.]

Most of these studies are somewhat fragmentary, and a detailed report of their findings is not given here. Instead, the test environment and results of the Aslib-Cranfield and SMART retrieval experiments, both of which include a large range of automatic text analysis methods, are described in some detail in the next section.

Retrieval System Evaluation

The work described above generally consists in implementing a particular type of text analysis process, and in testing it through use of a sample document collection and a set of sample queries. Both the Cranfield experiments, undertaken in England by Cleverdon and his associates, and the SMART project, based at Cornell and Harvard universities, have gone beyond that in the sense that a whole range of automatic text analysis methods were systematically tested, and that, at least in the case of SMART, the experimentation was extended to many different document collections in diverse fields, including documentation, computer engineering, aerodynamics, and medicine.

The Cranfield II experiments [not to be confused with the earlier Cranfield I tests designed to compare four conventional systems based on manual indexing (*28*)] were designed to measure a large variety of index-language "devices" that are potentially useful in the representation of document content. These devices include the use of synonym dictionaries, hierarchical subject classifications, phrase assignment methods, and many others. All the indexing tasks were performed manually by trained indexers, starting with the simple "single term" methods and proceeding to more complex methods involving use of a controlled vocabulary and various types of dictionaries. The indexing rules were carefully specified in each case, and were always based initially on the text of the documents or of the queries; the indexers were therefore simulating potential machine operations, and the evaluation results may thus be applicable to automatic indexing procedures.

A collection of 1400 documents in aerodynamics was available (the Cranfield collection) together with 279 search requests prepared by aerody-

Table 1. Order of effectiveness of three types of indexing languages. [Adapted from Cleverdon and Keen (*31*, fig. 8.1 T, p. 253)]

Type of indexing language	Rank orders for methods using indexing language	Average score for language
Single terms: content words manually chosen from full document	1, 2, 3, 4, 5, 6, 7, 12	64.15
Controlled terms: single terms modified by look-up in manually constructed thesaurus or authority list	10, 11, 15, 17, 18, 19	60.34
Simple concepts: single terms concatenated into standard noun phrases reflective of document content	8, 9, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33	54.55

namicists. Three main indexing languages were tested, known respectively as "single terms," "controlled terms," and "simple concepts." The single terms are content words chosen from document texts; controlled terms are single terms modified by "look-up" in a manually constructed subject authority list; and simple concepts are terms concatenated to form phrases. The test consisted in determining the retrieval effectiveness of these languages when used with the indexing devices referred to above. It was expected that some linguistic devices (the "recall devices"), including synonym dictionaries, concept associations, and term hierarchies, would broaden document and query identifications, thereby improving recall, while others (the "precision devices"), such as the assignment of term weights and the specification of relational indicators between terms, would narrow the content identifications, or make them more specific, thereby improving precision.

The evaluation process was based on a computation of recall and precision measures at various "coordination levels" (*29*)—that is, for various degrees of matching between queries and documents—followed by an averaging of results over all the search requests used. The output was then presented as a set of recall-precision tables and graphs. In addition, a global "normalized recall" measure, consisting, for each system, of a single value, computed in a manner somewhat analogous to computation of the "normalized" measures used in the SMART system (*30*), was used to rank the various systems in decreasing order of effectiveness. The detailed retrieval results (*31, 32*) cannot be reproduced here. However, a summary of the main results is contained in Table 1 where the three language types are arranged in decreasing order of effectiveness according to the average normalized recall score obtained.

It may be seen from Table 1 that the simple, uncontrolled indexing language involving single terms produces the best retrieval performance, while the controlled vocabulary and the phrases (simple concepts) furnish increasingly worse results. To quote from Cleverdon and Keen (*31*):

> . . . quite the most astonishing and seemingly inexplicable conclusion that arises from the project is that the single term index languages are superior to any other type . . .
> . . . of the six controlled term index language; . . . on the other hand, the single gave the best performance . . . as narrower, broader, or related terms are brought in, ranking orders . . . decrease . . .
> . . . the conceptual terms of the simple concept (phrase) index languages were over-specific when used in natural language; . . . on the other hand, 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 suffix cut-off process to generate word stems) could any improvement in performance be obtained. . . .

In other words, the surprising 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. Of the many procedures tried in an attempt to increase recall or precision, only the use of a synonym dictionary which groups related terms into concept classes produces a better performance than the original, unmodified terms. It goes without saying that single term indexing is easier to implement automatically than the more sophisticated, seemingly less effective alternatives.

One might be tempted to dismiss the Cranfield results by ascribing them to some peculiar test conditions were it not for the fact that the extensive evaluation work carried out for some years with the SMART system points in the same direction (*33, 34*). The SMART system is an experimental, fully automatic document retrieval system, operating with an IBM 7094 and a 360/65 computer. Unlike most other computer-based retrieval systems, the SMART system does not rely on manually assigned key words or index terms for the identification of documents and search requests, nor does it use primarily the frequency of occurrence of certain words or phrases included in the texts of documents. Instead, an attempt is made to go beyond simple word-matching procedures by using various intellectual aids in the form of synonym dictionaries, hierarchical arrangements of subject identifiers, statistical and syntactic phrase-generation methods, and the like, in order to obtain the content identifications useful for the retrieval process.

The following facilities incorporated into the SMART system for document analysis appear of principal interest.

1) A system for separating English words into stems and affixes (the so-called suffix "s" and stem thesaurus methods) used to construct document identifications consisting of the stems of words contained in the documents. Such a stem analysis is, preferably, applied only to those words whose frequency of occurrence in a given document is unexpectedly high, as compared with their frequency of occurrence in the literature at large (*10, 25*).

2) A synonym dictionary, or thesaurus, which can be used to recognize synonyms by replacing each word stem by one or more "concept" numbers; these concept numbers then serve as content identifiers, instead of the original word stem.

3) A hierarchical arrangement of the concepts included in the thesaurus which makes it possible, given any concept number, to find its "parents" in the hierarchy, its "sons," its "brothers," and any of a set of possible cross references. The hierarchy can be used to obtain more general content identifiers than the ones originally given (by going *up* in the hierarchy), more specific ones (by going *down*), and a set of related ones (by picking up brothers and cross-references) (*33*).

4) Statistical association procedures which use similarity coefficients based on term co-occurrences within the sentences of a document, or within the documents of a collection, to determine the "associated" terms. Such association methods then produce for each term a "profile" of associated terms, from which in turn a second-order profile containing still further associations

can be obtained, and so on (*35*); the original terms and their associations may then be used for content identification.

5) Syntactic analysis methods which make it possible to compare the syntactically analyzed sentences of documents and search requests. The syntactic analysis used to identify the phrases or sentence structures to be matched may be formal in the sense that it is based on a complete phrase structure or transformational grammar of the language, or the analysis may be of an ad hoc nature, based principally on recognition of certain function words from which prepositional and other phrases are then derivable (*36*).

6) Statistical phrase-matching methods which operate like the syntactic phrase procedures—that is, through use of a dictionary to identify phrases used as content identifiers. However, no syntactic analysis is performed in this case, and phrases are defined as equivalent if the phrase components match, regardless of the syntactic relationships between components.

7) A dictionary system, designed to revise the several dictionaries included in the system, such as the word stem dictionary, the word suffix dictionary, the common word dictionary (for terms to be deleted during analysis), the thesaurus (synonym dictionary), and the statistical and syntactic phrase dictionaries.

8) An automatic document classification system which groups documents with similar content identifiers into document clusters in such a way that a given file search can be confined to certain document clusters instead of being extended to the complete file.

9) A user feedback system which modifies document and query identifiers, on the basis of information supplied by the customers during the search process (*37*).

Stored documents and search requests are processed by the SMART system, without any prior manual analysis, by one of several dozen combinations of these and other automatic content analysis methods, and those documents which most nearly match a given search request are extracted from the document file in answer to each request.

In Table 2, a sample analysis produced by the SMART system, with a thesaurus process, is shown for query Q 13 B. The original query text is given, together with the resulting set of weighted (*38*) concept numbers (terms). Listed opposite each concept number is a sample of the terms appearing in the thesaurus under that concept category.

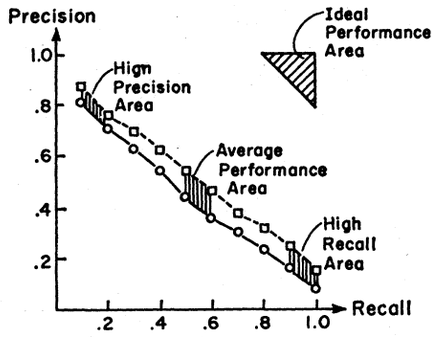

Fig. 2. Recall-precision graph illustrating word stem and thesaurus performance (averages for 780 Institute of Radio Engineers documents). (Circles) Word stem; (squares) thesaurus-3.

The SMART system organization makes it possible to evaluate the effectiveness of the various processing methods by comparing the output obtained from a number of different runs. This is achieved by processing the same search requests against the same document collections several times, while making selected changes in the analysis procedures between runs. By comparing the performance of the search requests under different processing conditions, it is possible to determine the relative effectiveness of the various analysis methods. The evaluation is then performed by averaging performance over many search requests and plotting recall-precision graphs of the type shown in Fig. 1. The effectiveness of a given method is reflected by the nearness of the corresponding curve to the upper right-hand corner of the graph, where both recall and precision are high.

Extensive evaluation results obtained with the SMART system have been published for collections in computer engineering, medicine, documentation, and aerodynamics (*33, 39, 40*). In each case, recall-precision graphs are drawn for the performance of two or more analysis and search procedures, averaged over many search requests, and the statistical significance of the differences in performance for any two methods is computed. A typical example, showing differences in performance for an antomatic word stem analysis and an analysis involving use of a stored synonym dictionary (or thesaurus) to transform weighted word stems into weighted thesaurus classes is shown in Fig. 2. It may be seen that, for the collection of 780 documents in computer engineering used with 35 search requests, the synonym recognition afforded by the thesaurus produces an increase in precision of about 10 percent for any given recall point.

It is not possible to reproduce here in detail the evaluation results obtained for many hundreds of runs. A few quotations from the published conclusions (slightly paraphrased to avoid introducing new terms not otherwise needed) may suffice (*40*, pp. 33–34).

> The order of merit is generally the same for all three collections [that is, computer engineering, aerodynamics, and documentation].
>
> The use of unweighted terms [weights restricted to 1 for terms that are present, and 0 for those that are absent] is always 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 always performs more effectively than the word stem methods where synonyms and other word relations are not recognized.
>
> The thesaurus and statistical phrase methods are substantially equivalent in performance; other dictionaries, including term hierarchies and syntactic phrases, perform less well.

These results indicate that, in automatic content analysis systems, weighted terms should be used, derived from document excerpts whose length is at least equivalent to that of a document abstract. Furthermore, synonym dictionaries should be incorporated wherever they are available. The principal conclusions reached by the Cranfield project are also borne out by the SMART studies: that phrase languages are not substantially superior to single terms as indexing devices, and that sophisticated analysis tools are less effective than had been expected.

Table 2. Thesaurus analysis for English query Q 13 B: "In what ways are computer systems being applied to research in the field of the belles lettres? Has machine analysis of language proved useful, for instance, in determining probable authorship of anonymous works or in compiling concordances?"

Concept Nos.	Weight	Sample terms in thesaurus category
3	12	Computer, processor
19	12	Automatic, semiautomatic
33	12	Analyze, analysis, etc.
49	12	Compendium, compile
65	12	Authorship, originator
147	12	Discourse, language
207	12	Area, branch, field
267	12	Concordance, KWIC†
345	12	Bell
*		Anonymous, lettres

*Query terms not found in thesaurus. †KWIC, "Key Word in Context."

Comparison of Automatic and Manual Indexing

The evaluation results described in the preceding section appear to raise as many questions as they answer. (i) What is the explanation for the finding, the reverse of what intuition leads one to expect, that simple automatic term extraction, combined with weighting and dictionary "look-up" methods, apparently produces a higher retrieval effectiveness than more sophisticated, semantically more complete, content analysis procedures such as complete word recognition, identification of pronoun referents, and analysis across sentence boundaries? (ii) How do the simple automatic indexing methods compare with conventional methods based on manual term assignment? (iii) How can the automatic procedures be improved (given that the performance range exemplified by the output of Fig. 2 is not as high as one would hope)? (iv) How would the automatic indexing process cope with the practical problems of automatic document input and of foreign language processing? (v) What is likely to be the future of automatic document processing? These questions are now treated in order.

The problem of rationalizing research results different from those one intuitively expects is always a difficult one. In the present case, however, some reasonable arguments are readily available.

First, it must be remembered that the problem of automatic documentation is not comparable to automatic translation or to automatic question answering, in that a retrieval system is designed only to lead a user to items likely to be related to the subject in which he is interested. A somewhat gross rendition of document content, consisting mostly of the more salient features, may therefore be perfectly adequate, in place of the line-by-line type of analysis needed, for example, for translation.

Second, a retrieval system is designed to serve a large, sometimes heterogeneous user population. Since users may have different needs and aims, and since their search requests may range from survey or tutorial type questions to very detailed analytical queries, an excessively specific analysis may be too specialized for most users.

Finally, in the evaluation procedures used to judge retrieval effectiveness, a performance criterion averaged over many search requests is used. This implies that analysis methods whose overall performance is moderately successful are given preference 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, a specific type of sophisticated analysis will be optimal, whereas, for the average query, the simpler type of indexing is best.

In explaining the test results, one might also argue that the evaluation results are inherently untrustworthy, first because they were obtained with small collections, often outside an accepted user environment and, second, because the recall and precision results are unreliable since they are based on subjective judgments of the relevance of the documents to the queries. Concerning the first point, it can be said that, although the tests were in fact conducted with collections of small size (less than 1500 documents each), the evaluation results are remarkably consistent over many collections in diverse subject areas; furthermore, the total test environment has included several thousand documents and several hundred queries. There is therefore no likelihood that such consistent results could have occurred by chance.

The second point appears, on the surface, more serious. It is a fact that recall and precision measures require a prior determination of relevance; that is, for each query it is necessary first to identify the set of relevant and nonrelevant items before the evaluation measures can be generated. Relevance assessments must be made by human subjects—preferably by the requester himself—and they will vary from one assessor to another. Studies of the relevance assessment process have indicated that the overall agreement between assessors may not be greater than about 30 percent (*41, 42*). Nevertheless, the conclusion that the recall and precision values are therefore unreliable is unwarranted. In fact, a recent study performed with four different sets of relevance assessments and a collection of 1200 documents in library science has shown that the average recall and precision curves are almost identical, even though the relevance sets are completely dissimilar. The explanation is that, for those documents which are most similar to the queries and which are therefore retrieved early in the search, the assessments are in almost perfect agreement; these documents are also the ones which principally determine the shape of the recall-precision curves in the nonzero regions, and which are therefore responsible for the relative invariance of the test results (*42*).

It appears, then, that reasonable arguments can be furnished to support the principal test conclusions, and that

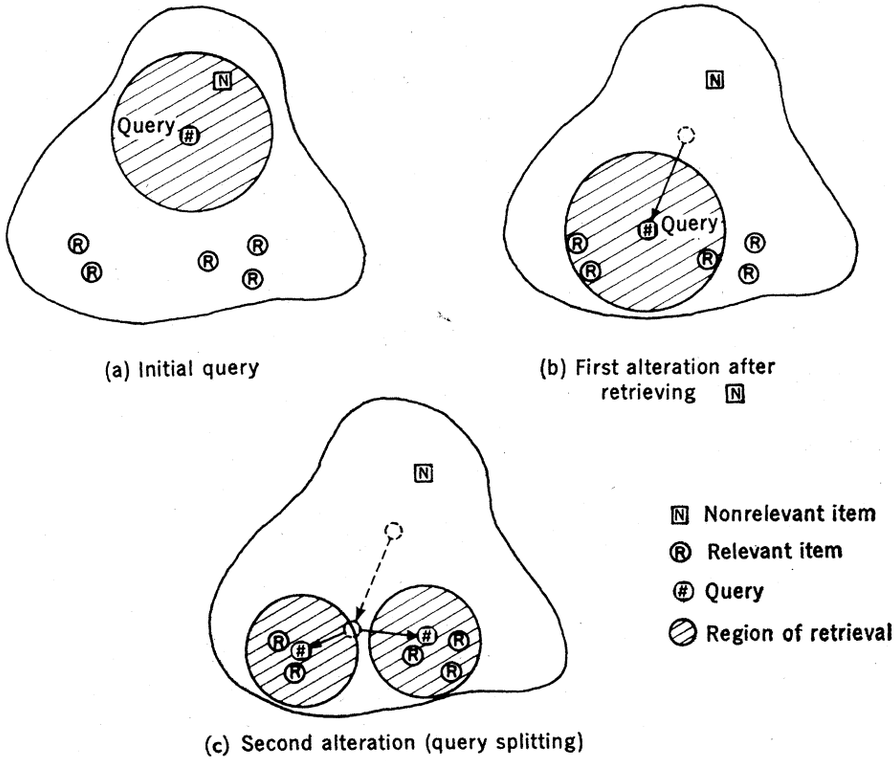

Fig. 3. Typical query transformations resulting from relevance feedback.

appropriate answers can be made in response to the more obvious objections. Finally, one must examine any counterevidence that may be available. Although systematic tests of automatic indexing procedures have not been made outside of the SMART and Cranfield environments, some data are available which appear not to be in agreement with the results reported above. For example, Saracevic reports that, in a test involving the use of 2600 documents in biomedicine and 124 queries, use of a thesaurus for term expansion was found not to be effective (*43*). It is not clear whether the fault, in this case, lay with the type of thesaurus [the SMART results apply only to certain types of thesauri, constructed in accordance with a specific set of principles (*33, 40*)] or with the type of analysis (a different analysis process was used for documents on the one hand and for queries on the other). Furthermore, because the results were cumulated for five different analysis procedures instead of being individually displayed, the output is not strictly comparable to the SMART or Cranfield data, and the results are difficult to assess.

The same is true of the test results obtained by Jones and his associates (*12*) using 22 queries each specified by a single phrase (or "content-bearing unit"). Here, a very high search precision (0.84) is reported for the phrase-matching process, but no recall values are given; the cited performance may thus correspond to a system operating at the left-hand end of a normal recall-precision curve. Furthermore, the queries, each consisting of a single two-word phrase, are probably not typical of queries normally received in information centers, and are in any case not comparable to the natural-language queries processed by SMART and Cranfield.

To summarize, there is no obvious evidence for distrusting the main results of the automatic indexing studies outlined above.

In some of the early text processing experiments it was seen that the automatic document search procedures were producing retrieval results at least equivalent to those obtained with conventional manual indexing (*20, 21*). Furthermore, the later tests conducted in an automatic retrieval environment indicate that the simple, single-term methods, which are easiest to implement on a computer, are also the most effective. It is interesting to try to determine under these circumstances how

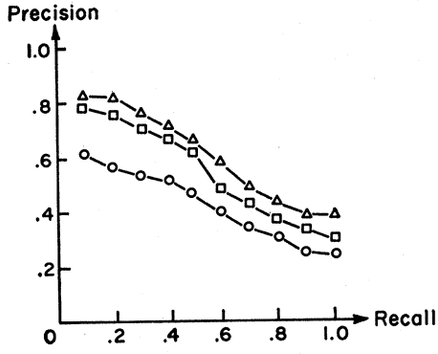

Fig. 4. Recall-precision graph illustrating the performance of abstract display and relevance feedback. (Circles) Original queries (word stem); (triangles) relevance feedback (word stem), one iteration; (squares) abstract display (word stem).

the automatic SMART procedures compare with standard manual indexing methods. The evidence here is not wholly conclusive, since the SMART processing is necessarily performed with small document collections. However, whatever evidence exists shows that the automatic indexing procedures are not inferior to what is now achieved by conventional, manual means.

For example, an initial comparison between the manual indexing used at Cranfield and the automatic processing of abstracts performed by SMART shows that the results obtained by the two systems are not statistically different (*31, 40*). To check these results, a comparison was made between the test results obtained by Lancaster for the manually based Medlars system (*17*) and the results for the SMART system. Specifically, for 18 of the Medlars queries used earlier by Lancaster, document abstracts were keypunched, and the retrieval process was repeated by means of the automatic text searching methods incorporated in SMART (*44*). The results indicate that, for that subcollection, a slightly higher average recall is obtained by SMART (0.69 as compared with 0.64 for Medlars), whereas Medlars achieves a somewhat higher precision (0.61 for SMART and 0.62 for Medlars). In any case, the intuitive feeling that the conventional indexing would necessarily be superior is again not confirmed.

One might interpret the results of the SMART-Medlars comparison by saying that the conventional and the automatic indexing procedures produce equally poor results (a recall and precision performance between 0.55 and 0.65, as compared with a possible maximum of 1). The reasons for the relatively poor performance of the automatic methods are clear when one considers the simplicity of the content analysis procedures used. For the manual indexing process, Lancaster reports the following main sources of failure (*17*): (i) index language problems (lack of specific terms or false coordination of terms); (ii) search formulation (query formulation too exhaustive or too specific); (iii) document indexing (document indexing insufficiently exhaustive, or too exhaustive, or important terms omitted); and (iv) lack of user-system interaction during the search process. The first three sources of failure all have to do with the query or document indexing process. The last inadequacy, however, appears to be one that can be remedied immediately.

For this reason, interactive search procedures have been incorporated into several recently implemented retrieval systems. The SMART system, in particular, attempts to meet the user problem by performing multiple rather than single searches. Thus, instead of submitting a search request and obtaining in return a final set of relevant items, a partial search is first made and, on the basis of the preliminary output obtained, the search parameters are adjusted before a second, more refined search is attempted. The adjustments made may differ from user to user, depending on individual needs, and the search process may be repeated as often as desired.

Various strategies are available for improving the results of a search by means of user feedback procedures (*37, 45, 46*). The first is based on a selective printout of stored information to be brought to the user's attention during the search process. For example, a set of additional, possible search terms related to those initially used by the requester may be extracted from the stored dictionary and presented to the user. The user may then be asked to reformulate the original query after selecting those new associated terms which appear to him most helpful in improving the search results. Typically, the statistical term associations discussed above can be used to obtain the set of related terms, or the sets of associated thesaurus classes can be taken from the thesaurus. This "search optimization" procedure is straightforward, but leaves the burden of rephrasing the query to the user (*45*).

A second strategy consists in automatically modifying a search request by using the partial results from a previous search. The user is asked to examine the documents retrieved by an initial

search, and to designate some of them as either relevant (R) or irrelevant (N) to his needs. Concepts from the documents termed relevant can then be added to the original search request if they are not already present, or, if they are, their importance can be increased through a suitable adjustment of weights; contrariwise, terms from documents designated irrelevant can be deleted or given a lower weight (*37*, *45–47*). An illustration of such a "relevance feedback" process is shown in Fig. 3, where a query first retrieves a document identified as nonrelevant (Fig. 3a). The query-updating process which follows then shifts the query in such a way that

230	ART	ARCHITEKTUR
231	INDEPEND	SELBSTAENDIG UNABHAENGIG
232	ASSOCIATIVE	
233	DIVIDE	
234	ACTIVE	AKTIV
	ACTIVITY	AKTIVITAET
	USAGE	TAETIGKEIT
235	CATHODE	DIODE
	CRT	VERZWEIGER
	DIODE	
	FLYING–SPOT	
	RAY	
	RELAIS	
	RELAY	
	SCANNER	
	TUBE	
236	REDUNDANCY	
	REDUNDANT	
237	CHARGE	EINGANG
	ENTER	EINGEGANGEN
	ENTRY	EINGEGEBEN
	INSERT	EINSATZ
	POST	EINSTELLEN
		EINTRAGUNG
238	MULTI–LEVEL	
	MULTILEVEL	
239	INTELLECT	GEISTIG
	INTELLECTUAL	
	INTELLIG	
	MENTAL	
	MIND	
	NON–INTELLECTUAL	
240	ACTUAL	PRAXIS
	PRACTICE	
	REAL	

Fig. 5. Excerpt from multilingual English-German thesaurus.

a new search operation retrieves some relevant documents (Fig. 3b). These documents, in turn, are used to generate two subqueries, which are then successful in retrieving all relevant items (Fig. 3c).

A good deal of work has been done to improve this type of feedback operation, and evaluation results indicate that the process is considerably more effective than the standard one-pass search process. Figure 4 is a typical feedback evaluation graph showing averages for 200 documents and 42 queries in aerodynamics. Here an initial one-step search process based on a word stem analysis is compared with a feedback procedure based on the display of abstracts of previously retrieved documents; such a display is then used for manual updating of queries. The results of such manual updating are in turn compared with one iteration of the automatic relevance feedback process. It may be seen in Fig. 4 that the automatic updating procedure is more effective than the manual one, and that an improvement of about 20 percent in precision is obtained through the feedback procedure. Moreover, this type of improvement in retrieval effectiveness has been duplicated for all collections so far processed (*45–47*).

Two other practical points—document input and foreign language processing—require discussion, since it is sometimes claimed that no automatic indexing process would be viable without consideration of these questions. The input problem is particularly acute in an environment which includes automatic indexing, since document excerpts of at least abstract length should be available for analysis. Obviously, if all the material contained in an abstract (or, possibly, in the full text) requires manual keypunching, the main benefits of the automatic analysis procedure may be lost. No overall solution appears immediately available. However, the use of automatic character recognition equipment and of automatic typesetting processes is becoming more widespread, with the result that document input products that can be automatically read may well become generally available with each document before long.

Concerning the foreign language problem, the situation is less difficult than might appear to be the case. It is true that, in certain subject areas, up to 50 percent of the pertinent documents are not written in English [this is true of the documents in biomedicine processed at the National Library of Medicine (*17*)]. The English-language analysis methods will obviously not be applicable for these documents. However, it is also true that 90 percent of these documents are in one of only six or seven languages, most of them being in French, German, and Russian.

Table 3. Thesaurus analysis for German query Q 13 B: *"Inwieweit werden Computersysteme zur Forschung auf dem Gebiet der schönen Literatur verwendet? Hat sich maschinelle Sprachenanalyse als hilfreich erwiesen, um z. B. die vermutliche Autorenschaft bei anonymen Werken zu bestimmen oder um Konkordanzen zusammenzustellen?"**

Concept Nos.	Weight	Sample terms in thesaurus category
19†	12	Computer, Datenverarbeitung
3†	12	Automatisch, Kybernetik
21	4	Artikel, Presse, Zeitschrift
33†	6	Analyse, Sprachenanalyse
45	4	Herausgabe, Publikation
64	4	Buch, Heft, Werk
65†	12	Autor, Verfasser
68	12	Literatur
147†	6	Linguistik, Sprache
207†	12	Arbeitsgebiet, Fach
267†	12	Konkordanz, KWIC
‡		schönen, hilfreich, vermutlich, anonymen, zusammenzustellen

*For translation, see Table 2. †Common concepts with English query. ‡Query terms not found in thesaurus.

Some experiments were recently conducted with the SMART system and a collection of about 500 German documents in the field of library science. A multilingual thesaurus (Fig. 5) was prepared manually by translating the English version of an existing thesaurus into German. From Fig. 5 it may be seen that the same concept-class number represents both an English word class and the corresponding German class. The translation test performed consisted in processing a set of original English language queries against both the English and the German document collections; the test was then repeated by processing the English queries manually translated into German against the same two collections (English and German). The test results indicate that no significant loss in performance results from the process of query translation (*48*).

A sample German query processed through the German thesaurus is shown in Table 3. A comparison with Table 2 shows that a large number of "English" concepts are also present in the German analysis, and this accounts for the fact that the thesaurus translation is successful. The foreign language problem appears not to present a major roadblock to development of an automatic document processing system.

Summary

A large number of automatic text analysis and indexing experiments have been examined. All the available evidence indicates that the presently known text analysis procedures are at least as effective as more conventional manual indexing methods. Furthermore, a simple indexing process based on the assignment of weighted terms to documents and search requests produces better retrieval results than a more sophisticated content analysis based on syntactic analysis or hierarchical term expansion. Such a simple automatic indexing procedure is easily implemented on present-day computers, and there are no obvious technical reasons why manual document analysis methods should not be replaced by automatic ones.

While automatic document analyses appear, therefore, to be at least as efficient as presently used manual methods, it is unfortunately the case that all known indexing procedures—whether manual or automatic—produce relatively mediocre results. One of the most fruitful ways of upgrading retrieval performance consists in using multiple searches based on user feedback information furnished during the search process. Interactive search methods should then lead to a retrieval effectiveness approaching a recall and precision of about 0.70 instead of the present 0.50 to 0.60. Some tentative extrapolations appear to indicate that an increased sophistication in indexing and search methodology may eventually lead to "optimal" systems for which the average recall and precision values would approach 0.80 (*42*, *49*).

No obvious advances leading to additional large-scale improvements in retrieval effectiveness are likely to be made soon. For this reason, the known automatic document analysis and search procedures described in this article may well become the standard tools in most mechanized information systems of the future.

References and Notes

1. Y. Bar Hillel, in *Digitale Informationswandler*, W. Hoffman, Ed. (Vieweg, Brunswick, Germany, 1962).
2. H. P. Luhn, *IBM J. Res. Develop.* **1**, No. 4 1957).
3. ———, "Potentialities of Auto-Encoding of Scientific Literature," *IBM Res. Cent. Rep. No. RC-101* (1959).
4. J. O'Connor, *J. Ass. Comput. Mach.* **11**, 437 (1964); M. E. Stevens, "Automatic Indexing: A State of the Art Report," *U.S. Nat. Bur. Stand. Monogr. 91* (1965); ———, V. E. Giuliano, L. B. Heilprin, Eds., "Statistical Association Methods for Mechanized Documentation," *U.S. Nat. Bur. Stand. Monogr. 269* (1965).
5. C. Montgomery and D. R. Swanson, *Amer. Doc.* **13**, 359 (1962).
6. M. J. Ruhl, *Amer. Doc.* **15**, 136 (1964).
7. D. H. Kraft, *Amer. Doc.* **15**, 48 (1964).
8. J. O'Connor, *Amer. Doc.* **15**, 96 (1964).
9. H. Fangmeyer and G. Lustig, "The Euratom automatic indexing project," paper presented before IFIP [International Federation for Information Processing] Congress 68, Edinburgh (1968).
10. F. J. Damerau, "An Experiment in Automatic Indexing," *IBM. Res. Cent. Rep. No. RC-894* (1963).
11. M. E. Stevens and G. H. Urban, in *Proceedings, Spring Joint Computer Conference* (Spartan, Washington, D.C., 1964), pp. 563–575; T. N. Shaw and H. Rothman, *J. Doc.* **24**, No. 3 (1968).
12. P. E. Jones, V. E. Giuliano, R. M. Curtice, "Papers on Automatic Language Processing—Development of String Indexing Techniques," *Arthur D. Little (Cambridge, Mass.) Rep. ESD-TR-67-202* (1967), vol. 3.
13. M. M. Kessler, *Amer. Doc.* **16**, 223 (1965).
14. R. M. Needham, *Mech. Transl.* **8**, Nos. 3 and 4 (June–Oct. 1965); L. B. Doyle, *J. Ass. Comput. Mach.* **12**, No. 4 (1965); H. Borko and M. D. Bernick, *J. Ass. Comput. Mach.* **10**, No. 2 (1963).
15. M. E. Maron, *J. Ass. Comput. Mach.* **8**, 404 (1961).
16. J. O'Connor, *J. Ass. Comput. Mach.* **12**, 490 (1965).
17. F. W. Lancaster, "Evaluation of the Operating Efficiency of Medlars," *Nat. Libr. Med. Final Rep.* (1968).
18. These figures imply that an average search processed by Medlars manages to retrieve almost 60 percent of what is wanted, while only half the retrieved items are not relevant; in view of the large document file being processed—over 600,000 items—this is a remarkable achievement.
19. P. Atherton, D. W. King, R. R. Freeman, "Evaluation of the Retrieval of Nuclear Science Document References Using UDC as the Indexing Language for a Computer Based System," *Amer. Inst. Phys. Rep. AIP-UDC 8* (1968); F. H. Barker and D. C. Veal, "The Evaluation of a Current Awareness Service for Chemists," *Chem. Soc. Res. Unit Inform. Dissemination Retrieval Rep.* (1968); C. D. Gull, in *Coordinate Indexing*, M. Taube *et al.*, Eds. (Documentation, Inc., Washington, D.C., 1963); L. B. Heilprin and S. S. Crutchfield, in *Proceedings, 1964 Annual Meeting of ADI [American Documentation Institute]* (Spartan, Washington, D.C., 1964); M. R. Hyslop, *ibid.*; W. F. Johanningsmeier and F. W. Lancaster, "Project SHARP Information Storage and Retrieval System: Evaluation of Indexing Procedures and Retrieval Effectiveness," *U.S. Bur. Ships Rep. NAVSHIPS 250-210-3* (1964); D. W. King, *J. Chem. Doc.* **5**, 96 (1965); D. B. McCarn and C. R. Stein, in *Electronic Handling of Information: Testing and Evaluation*, A. Kent *et al.*, Eds. (Thompson, Washington, D.C., 1967), pp. 110-122; E. Miller, D. Ballard, J. Kingston, M. Taube, in *Proceedings, International Conference on Scientific Information* (National Academy of Sciences–National Research Council, Washington, D.C., 1959), vol. 1, pp. 671–685; B. A. Montague, *Amer. Doc.* **16**, 201 (1965); National Academy of Sciences Ad-Hoc Committee of the Office of Documentation, "The Metallurgical Searching Service of the American Society of Metals—Western Reserve University: An Evaluation," *Nat. Acad. Sci. Nat. Res. Counc. Publ. 1148* (1964); J. A. Schuller, *Aslib Proc.* **12**, 372 (1960); J. Tague, *Effectiveness of a Pilot Information Service for Educational Research Materials* (Center for Documentation and Communication Research, Western Reserve University, Cleveland, 1963).
20. D. R. Swanson, *Science* **132**, 1099 (1960).
21. ———, in *Proceedings IFIP [International Federation for Information Processing] Congress 62*, C. Popplewell, Ed. (North Holland, Amsterdam, 1963), pp. 288–293.
22. E. M. Fels, *Amer. Doc.* **14**, No. 1 (1963).
23. B. Altmann, *Amer. Doc.* **18**, No. 1 (1967).
24. J. S. Melton, "Automatic Processing of Metallurgical Abstracts for the Purpose of Information Retrieval," *Center for Communication and Documentation Research, Case Western Reserve University, Cleveland, Final Rep. NSF-4* (1967).
25. S. F. Dennis, in *Information Retrieval—a Critical View*, G. Schechter, Ed. (Thompson, Washington, D.C., 1967).
26. V. E. Giuliano and P. E. Jones, "Study and Test of a Methodology for Laboratory Evaluation of Message Retrieval Systems," *Arthur D. Little [Cambridge, Mass.] Rep. ESD-TR-66-405* (1966).
27. K. Sparck Jones and D. M. Jackson, "The Use of Automatically Obtained Keyword Classifications for Information Retrieval," *Cambridge [England] Language Research Unit Final Rep. ML 211* (1969).
28. C. W. Cleverdon, in *Proceedings, International Conference on Scientific Information* (National Academy of Sciences–National Research Council, Washington, D.C., 1959), vol. 1, pp. 687–698; ———, "Testing and Analysis of an Investigation into the Comparative Efficiency of Indexing Systems," *Cranfield (England) Res. Proj. Rep.* (1962).
29. A document exhibiting n terms in common with a given query is said to be retrieved at coordination level n.
30. G. Salton, in *Proceedings, FID Congress 1965* (Spartan, Washington, D.C., 1966).
31. C. W. Cleverdon and E. M. Keen, "Factors Determining the Performance of Indexing Systems," *Aslib Cranfield (England) Res. Proj. Rep.* (1966), vols. 1 and 2.
32. C. W. Cleverdon, *Aslib Proc.* **19**, No. 6 (1967).
33. G. Salton, *Automatic Information Organization and Retrieval* (McGraw-Hill, New York, 1968).
34. ——— and M. E. Lesk, *Commun. Ass. Comput. Mach.* **8**, No. 6 (1965); G. Salton, *IEEE (Inst. Elec. Electron. Eng.) Spectrum* **2**, No. 8 (1965).
35. H. E. Stiles, paper presented at the 3rd Institute on Information Storage and Retrieval, American University, Washington, D.C., 1961.
36. P. B. Baxendale, *IBM J. Res. Develop.* **4**, No. 2 (1958); D. J. Hillman and A. J. Kasarda, in *Proceedings AFIPS Spring Joint Computer Conference* (AFIPS Press, Montvale, N.J., 1969).
37. G. Salton, in *Proceedings IFIP [International Federation for Information Processing] Congress 68* (North-Holland, Amsterdam, 1969).
38. In the SMART system the term weights are based on the frequency of occurrence of each term in a document, as well as on individual term charateristics derived from the word stem or from thesaurus dictionaries.
39. G. Salton, *Amer. Doc.* **16**, 209 (1965); ——— *et al.*, "Information Storage and Retrieval," *Dep. Computer Sci. Cornell Univ. Rep. Nos. ISR-11, ISR-12, ISR-13, ISR-14, and ISR-16 to Nat. Sci. Found.* (1966–1969).
40. G. Salton and M. E. Lesk, *J. Ass. Comput. Mach.* **15**, 8 (1968).
41. C. A. Cuadra and R. V. Katter, *J. Doc.* **23**, No. 4 (1967); A. M. Rees and D. G. Schultz, "A Field Experimental Approach to the Study of Relevance Assessments in Relation to Document Searching," *Cent. Doc. Commun. Res. Case Western Reserve Univ. Final Rep. to Nat. Sci. Found.* (1967).
42. M. E. Lesk and G. Salton, *Inform. Storage Retrieval* **4**, No. 4 (1968).
43. T. Saracevic, "The Effect of Question Analysis and Searching Strategy in Performance of Retrieval Systems: Selected Results from an Experimental Study," *Comp. Syst. Lab. Rep. No. CSL:TR:5, Cent. Doc. Commun. Res. Case Western Reserve Univ. Cleveland* (1968).
44. G. Salton, *Amer. Doc.* **20**, No. 1 (1969).
45. M. E. Lesk and G. Salton, in *Proceedings AFIPS Spring Joint Computer Conference* (AFIPS Press, Montvale, N.J., 1969).
46. E. Ide, "Relevance Feedback in an Automatic Document Retrieval System," *Dep. Computer Sci. Cornell Univ. Rep. No. ISR-15 to Nat. Sci. Found.* (1969).
47. G. Salton, in *Mechanized Information Storage, Retrieval and Dissemination*, K. Samuelson, Ed. (North-Holland, Amsterdam, 1968), pp. 73–107.
48. ———, "Automatic processing of foreign language documents," *J. ASIS [Amer. Soc. Inf. Sci.]*, in press.
49. C. W. Cleverdon, "The Methodology of Evaluation of Operational Information Retrieval Systems based on the Test of Medlars," *Cranfield (England) Res. Proj. Rep.* (1968).

50. This study was supported in part by the National Science Foundation under grant GN-750.