Automatic Text Analysis

G. Salton

Technical Report
No. 69-36
June 1969

Department of Computer Science
Cornell University, Ithaca, N. Y.

Automatic Text Analysis

G. Salton*

Abstract

Effective automatic methods are now available to replace conventional document indexing and classification.

1. Introduction

> "...it seems, therefore, that the only steps of the literature search process which are amenable to performance by a digital computer are those steps which follow the assignment of the Boolean function over the topic terms, up to and including the printing of the reference list." [1]

These words, written about ten years ago, reflect the writers conviction that the role of computers in information retrieval would be confined to a simple matching operation between term sets manually assigned to documents, and to information requests, followed by the withdrawal from the file of certain document citations whose terms are similar to the query terms. This excludes in particular all types of "automatic indexing" and "automatic classification" operations of the kind used automatically to assign terms, or content identifiers, to documents and search requests, and to group the items automatically into certain subject classes.

---

*Professor of Computer Science, Cornell University, Ithaca, N. Y. 14850

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

In the last ten years, dozens of experiments have however been conducted in the general area of automatic text processing, including not only automatic query-document matching, but also indexing and classification, as well as interactive procedures designed to produce better retrieval effectiveness for individual customers. Furthermore, there is now overwhelming evidence to show that the automatic content analysis and text processing methods are not only relatively easy to implement, but also produce a retrieval effectiveness at least equal to that obtained by the conventional, mostly manual procedures used in the past. As a result, it is no longer idle to speculate about fully-automatic document processing centers, and the automatic library may well be just around the corner.

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

2. General Methodology

The first serious work in automatic text analysis dates back to the middle and late nineteen fifties, 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:

> "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." [2, p. 315]

Luhn further suggested that the inquirer's "document" (that is, the search request) would 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 — too 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:

a) 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 — for example, the frequency of occurrence of the given expression should neither be too high nor too low, and the expression should not be included in a "negative" dictionary of prohibited terms.

b) a weight may be assigned to each expression based on the frequency of occurrence of the given expression, or on the position of the expression in the document, or on the type of entity.

c) 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 cooccurrence characteristics between the terms in a document collection, or on syntactical relations between words.

d) 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 content analysis purposes.

It is impossible in the present context to relate in detail the many strategies which have been proposed for automatic indexing. [4,5,6] Instead the automatic SMART document retrieval system which has been operating for some years on an IBM 7094 and the 360/65, and which includes most of the more common automatic content analysis procedures is used as an example. [7,8,9] The following facilities incorporated into the SMART system for document analysis appear of principal interest:

a) a system for separating English words into stems and affixes (the so-called suffix 's' and stem thesaurus

methods) which can be used to construct document identifications consisting of the stems of words contained in the documents; such a stem analysis can be applied to all the words in a text, except certain common function words, or preferably it might be applied only to those words whose frequency of occurrence in a given document is unexpectedly high, compared with their frequency of occurrence in the literature at large. [10,11]

b) 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 stems; such a thesaurus may be constructed manually by studying the vocabulary characteristics in a given subject area, and grouping related terms into concept categories [7], or, alternatively term groups may be generated automatically, by clustering techniques designed to group together all those terms which exhibit high cooccurrence characteristics in the documents of a collection. [12]

c) 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; a term hierarchy may again be constructed manually, or by automatic methods using term cooccurrence characteristics as a basis. [7]

d) statistical association procedures which use similarity coefficients based on term cooccurrences 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; [13] the original terms and their associations may then be used for content identification.

e) syntactic analysis methods which make it possible to compare the syntactically analyzed sentences of documents and search requests with a pre-coded dictionary of syntactic structures ("criterion trees") in such a way that the same concept number is assigned to a large number of semantically equivalent, but syntactically quite different constructions; 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 else, the analysis may be conducted on an ad-hoc basis by recognizing principally certain function words from which prepositional and other phrases are then derivable. [14,15]

f) statistical phrase matching methods which operate like the preceding syntactic phrase procedures, that is, by using a preconstructed 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 concept numbers of all components match, regardless of the syntactic relationships between components; various criteria can be used to decide on the acceptability of a given phrase as a content indicator: some of the principal ones are the coherence between phrase components, that is, the frequency of cooccurrence of the components over and above random expectation; the repeatability, that is, the critical frequency of occurrence of the phrase; the accountability, that is, the extent to which the occurrence of a component within a phrase accounts for a minimum percentage of the

total number of occurrences of the component; and the uncommonality of the phrase which rules out phrases which are strictly syntactic in nature. [16]

g) a dictionary system, designed to revise the several dictionaries included in the system:

   i) word stem dictionary
   ii) word suffix dictionary
   iii) common word dictionary (for words to be deleted during analysis)
   iv) thesaurus (synonym dictionary)
   v) concept hierarchy
   vi) statistical phrase dictionary
   vii) syntactic ("criterion") phrase dictionary.

h) 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 then be confined to certain document clusters only instead of being extended to the complete file.

i) a user feedback system which modifies document and query identifiers, based on information supplied by the customers during the search process; this insures that documents presumed to be relevant to certain queries are more easily retrievable by being placed in closer proximity to these queries.

A sample analysis produced by the SMART system using a thesaurus process is shown for query Q 13 B in Table 1. The original query text is given together with the resulting set of weighted concept numbers. Furthermore, each concept number is listed with a sample of the terms appearing in the thesaurus under that concept category. Following such a thesaurus analysis, retrieval would take place by matching the query "concept vector" with concept vectors derived from the document abstracts or texts, and retrieving matching items.

The experimental evidence derived from many of the studies in automatic text analysis is examined in the remainder of this report, and conclusions are drawn concerning the efficacy of the various techniques.

3. 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: the title word studies, the automatic versus manual term studies, and the 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 manually assigned subject headings. Montgomery and Swanson used an issue of Index Medicus containing title citations in biomedicine cross-filed under the various manually derived subject headings, and concluded that for 86% of almost 5000 titles a correlation existed between the subject heading assigned in Index Medicus and the document title (correlation being defined as an actual match between word stems, or a match between somewhat loosely defined synonymous terms). [17] In a somewhat related study using the chemical literature, Ruhl found that 57% of the titles

examined contained all important concepts (or their equivalents) listed for these documents in the subject index of Chemical Abstracts, while only 12% of the titles missed three or more important subject headings. [18] Similar results were found by Kraft for the legal literature, where only about ten percent of the document titles examined did not contain any keywords useful for indexing purposes, while 64% of the title entries contained one or more of the subject heading words included in the "Index to Legal Periodicals", and an additional 25% of the titles contained "logical equivalents" of the subject headings. [19]

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 completely worthless. Furthermore, the counterevidence cited by O'Connor, who finds a correlation between assigned subject headings and title words ranging from a low of only 13% to a high of only 68%, 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 optimal either for indexing or for retrieval. [20]

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 based on statistics derived from a manually indexed test collection which provides the conditional probabilities that index term A may be assigned to a document given that word B occurs a certain number of times in the document text. These conditional probabilities are then used for the automatic assignment

of terms to a new control collection, following a machine analysis of the document texts. Finally, these automatically generated index term sets are compared with the available manually assigned terms, using an evaluation coefficient such as q to measure the amount of overlap between vocabularies, where

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

here c represents the set of common term assignments, a is the set of automatically derived terms, and m is the set of manually assigned terms. [21]

Various tests of this general type have been performed [10,22 23], and the consensus is that about sixty percent agreement is obtainable between manually and automatically produced terms. In one particular test involving automatic phrase generation, using the previously mentioned criteria of coherence, repeatability, accountability and uncommonality, as many as 86% of the automatically assigned phrases were found to be acceptable subject heads by human judges, the overassignment being of the order of 14%, with an underassignment (that is, proper content indicators not recognized by machine) of the order of 11%. [16] A related approach, consisting of a comparison between automatically derived document correlations based on similar bibliographic citation patterns with document similarities based on overlapping sets of manually assigned subject headings also indicates a considerable amount of agreement between the automatic and manual procedures. [24]

The last set of experiments not including tests in a retrieval environment involves the automatic classification of documents into

subject categories (rather than the assignment of index terms to documents). [25,26,27] Here again, a test collection may be 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 new, incoming documents. [28,29] It is found that for the documents originally used to derive the test parameters, an automatic assignment to subject categories is about eighty to ninety percent effective (that is, the correct category is chosen in about eighty to ninety percent of the cases). For documents not used in deriving the test parameters, the effectiveness of the automatic classification based on document vocabularies drops down to about fifty percent.

O'Connor [29] remarks that the percentage of correctly classified documents increases when more refined classification parameters are used (from 76% when keywords alone are used to 92% when certain relationships between keywords 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%. 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 and will be noticed again in the experimental results to be reported in the remainder of this study.

4. Retrieval Experiments

The indexing experiments previously described were not performed within a normal retrieval situation, and relied on criteria supplied by

human subject experts for evaluation purposes. Over the last few years, the preferred way to test indexing, or classification, or search procedures has been to include them within a retrieval system, either experimental or real, and to judge the effectiveness of the various devices as a function of their performance as part of such a system.

The evaluation of retrieval systems can be carried out in many different ways, depending on the type of system considered — whether operational, experimental with user population, or laboratory type system; on the viewpoint taken — that of the user, the manager, or the operator; and on other factors. A large number of different variables may affect the results of any evaluation process, including the kind of user population, the type and coverage of the document collection, the indexing tools, the analysis and search methods incorporated into the system, the equipment used, the operating efficiency, as well as costs and time lag needed to produce answers, and many others.

In many of the principal evaluation studies, the viewpoint taken is the user's, and the overriding criterion of system effectiveness is the ability of the system to satisfy the user's information need 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, while, at the same time, 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 precision figures achievable by a given system are adjustable, in the sense that a relaxation of search condition (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, since the retrieval of more relevant items normally also leads to the retrieval of more irrelevant ones. When recall and precision are plotted against each other on a graph, a monotonically decreasing curve of the type shown in Fig. 1 thus 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 which are also retrieved is kept within tolerable limits. Thus, in what is probably the most exhaustive evaluation of an operating retrieval system using manually indexed documents (the Medlars system at the National Library of Medicine), Lancaster reports an average recall of 0.577 and an average precision of 0.504.* [30] Comparable data are available from the extensive literature dealing with the evaluation of operating, manually based retrieval systems. [31-43]

The first comparison of conventional retrieval, using manually

---

*This implies that an average search processed by Medlars manages to retrieve almost sixty 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.

indexed documents, with automatic text processing systems appears to be the one performed by Swanson in the late nineteen-fifties using one hundred documents and 50 queries. [44] Three indexing and analysis systems were used, including conventional retrieval based on a subject heading index, retrieval based on specifications provided by words and phrases automatically extracted from the document texts, and, finally, retrieval using a thesaurus in addition to the words obtained from the documents. A recall-precision like measure was used to evaluate system performance, which varied directly with the relevance weight of retrieved items, and included in addition a penalty factor for irrelevant material also retrieved.

The test results indicated that the average retrieval performance based on the automatic text analysis was superior to 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 the report: [44]

> "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 framework 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 to be reported later in this study, these prophesies must appear as remarkably accurate.

The original results of Swanson 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). [45] Documents were retrieved in decreasing order of similarity with the queries, the similarity score of 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 nth document); this produces a sequence of recall-precision pairs which can be plotted as a curve similar to the one shown in Fig. 1. Once again, Swanson concludes his study by stating: [45]

> "though these results (that automatic text processing using an automatic thesaurus is more accurate than the human processing 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".

Following these first two experiments, various other studies also included elements of automatic text analysis, including full text search [46], the use of phrase dictionaries and syntactic analysis procedures [47,48], statistical term associations [11,49], and automatically constructed term groupings (thesauruses) [12]. In each case, the intent is to show that one, or another of the proposed automatic language analysis methods operate more successfully than either a manual indexing process, or an automatic process using a less sophisticated approach. In general, the case is made that the use of manually constructed thesauruses, 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 [48]; furthermore, retrieval appears to have been based on the presence, or absence, of a single matching phrase, or sentence fragment so that the test results are difficult to interpret effectively.)

Since these somewhat fragmentary results are generally subsumed in the test environment of the Aslib-Cranfield and the SMART retrieval experiments, both of which include a large range of automatic text analysis methods, a detailed report of their findings is not made here.

5. Retrieval System Evaluation

The work described in the preceding section generally consists in implementing a particular type of text analysis process, and in

testing it using a sample document collection and a set of sample queries. Both the Cranfield experiments undertaken in England by Cleverdon and 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 for the SMART case, 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 [50,51]) were designed to measure a large variety of index language "devices" which are potentially useful in the representation of document content, including the use of synonym dictionaries, hierarchical arrangments of content identifiers, phrase assignment methods and many others. All the indexing tasks were performed manually by trained indexers, starting with the simple "single term" methods to the more complex procedures using a controlled vocabulary together with various types of dictionaries. The indexing rules were carefully specified in each case, and were always based initially on the document or query texts; 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 aerodynamicists. Three main indexing languages were tested, known respectively as single terms, controlled terms, and simple concepts —

where 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 earlier. The expectation was that some devices, including synonym dictionaries, concept associations, and term hierarchies would serve to broaden document and query identifications thereby improving recall (the "recall devices"), while others such as term weighting and use of relational indicators would narrow the content identifications, or make them more specific, thereby improving precision (the "precision devices").

The evaluation process was based on a computation of recall and precision measures at various "coordination levels", 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 the normalized measures used by the SMART system [52], was used to rank the various systems in decreasing order of effectiveness. The detailed retrieval results cannot be reproduced here. [53,54] However, a summary of the main results is contained in Table 2, where the three language types are arranged in decreasing order according to the average normalized recall score obtained.

It is seen from Table 2 that the simple, uncontrolled indexing

language using single terms produces the best retrieval performance, while the controlled vocabulary and the phrases (simple concepts) furnishes increasingly worse results. To quote from Cleverdon: [53]

> "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 (p. 252)..."
>
> "of the six controlled term index languages, that using only the basic terms gave the best performance... as narrower, broader, or related terms are brought in, ranking orders... decrease (p. 254)..."
>
> "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 (p. 255)...".

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 possibly weighted terms obtained from document or query texts are also the most effective. Of the many recall and precision devices tried, 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 much 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, if it were not for the fact that the extensive evaluation work carried out for some years with the SMART system point in the same direction. [7,8,9] The SMART system is an experimental, fully-automatic document retrieval system, operating on an IBM 7094 and a 360/65 computer. Unlike 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 a variety of 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.

Stored documents and search requests are then processed with any prior manual analysis by one of several hundred automatic content analysis methods, and those documents which most nearly match a given search request are extracted from the document file in answer to the request. The system may be controlled by the user, in that a search request can be processed first in a standard mode; the user can then analyze the output obtained and, depending on his further requirements order a reprocessing of the request under new conditions. The new output can again be examined and the process iterated until the right kind and amount of information are retrieved. [55] SMART is thus designed as an experimental automatic retrieval system of the kind that may become current in operational environments some years hence.

The SMART system organization makes it possible to evaluate the effectiveness of the various processing methods by comparing the output obtained from a variety 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 then possible to determine the relative effectiveness of the various analysis methods. The evaluation is actually 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 then 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. [7,56,57,58] In each case, recall-precision graphs are drawn for two or more analysis and search procedures, averaged over many search requests, and the statistical significance of the differences in performance between any two methods is computed. A typical example, showing differences between an automatic word stem analysis, and an analysis using a stored synonym dictionary (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 improvement of about ten percent in precision 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: [57, p. 33-34]

a) the order of merit is generally the same for all three collections (that is, computer engineering, aerodynamics, and documentation);

b) the use of logical vectors (that is, term vectors in which term weights are disregarded) is always less effective than the use of weighted terms;

c) the use of document titles alone is always less effective for content analysis purposes than the use of full document abstracts;

d) the thesaurus process involving synonym recognition always performs more effectively than the word stem methods where synonyms and other word relations are not recognized;

e) the thesaurus and statistical phrase methods are substantially equivalent in performance; other dictionaries, including term hierarchies and syntactic phrases perform less well.

These results thus 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 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 so-

phisticated analysis tools — other than simple synonym recognition — are not as effective as expected.

6. Automatic versus Manual Indexing

The evaluation results described in the preceding section appear to raise as many questions as they answer: first, what is the explanation for the wholly counterintuitive notion that simple automatic term extraction, combined with weighting and dictionary look-up methods apparently produce a higher retrieval effectiveness than more sophisticated, semantically more complete content analysis procedures; second, how do the simple automatic indexing methods compare with conventional methods based on manual term assignment; third, 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; fourth, how would the automatic indexing process cope with the practical problems of automatic document input and of foreign language processing; and fifth, what is likely to be the future of automatic document processing. These questions are now treated in order.

A) Reliability of Indexing Results

The problem of rationalizing research results which are not intuitively expected 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 the sense that a retrieval system is designed only to lead a user to items likely to be related to his interest; a somewhat gross rendi-

tion of document content consisting mostly of the more salient features may therefore be perfectly adequate, instead of a line-by-line type of analysis needed, for example, for translation purposes.

Second, a retrieval system is designed to service a large, sometimes heterogeneous user population; that implies that facilities must be available to help the average user. Since users may have different needs and aims, ranging all the way from survey or tutorial type questions to very detailed analytical queries, an excessively specific analysis may be too specialized for most users. This is reflected in Cleverdon's conjecture that there exists a correct level of specificity for the analysis of each document; if this level is too high, the average performance degrades.

Finally, the evaluation procedures used to judge retrieval effectiveness utilize a performance criterion averaged over many search requests. This implies that preference is given to analysis methods whose overall performance is moderately successful, over possibly more sophisticated procedures which may operate excellently for certain queries but much less well for others. In practice, it may then 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 relevance judgments of

the documents with respect 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 more serious on the surface. 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 requestor 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 thirty percent. [59,60,61] 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, there exists in fact almost perfect agreement in the assessments; these documents are, however, 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. [61]

It appears then that reasonable arguments can be furnished to support the principal test conclusions, and that appropriate answers are available to respond to the more obvious objections. There remains then to examine any counterevidence that might 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 earlier. For example, Saracevic reports that in a test using 2600 documents in biomedicine together with 124 queries, a thesaurus used for term expansion was not found to be effective. [62] It is not clear in this case whether the thesaurus is at fault — the SMART results apply only to certain types of thesauruses, constructed in accordance with a specific set of thesaurus construction principles [7,57] — or the type of analysis — a different analysis process was used for documents on the one hand and for queries on the other, and the results were cumulated for five analysis procedures instead of being individually displayed. In any case, the output is not strictly comparable with the SMART or Cranfield data, and the results are difficult to assess.

The same is true of the test results obtained by Jones and associates using 22 queries each specified by a single phrase (or "content-bearing unit"). [16] Here, a very high search precision is reported for the phrase matching process (0.84) but no recall values

are given; the cited performance may then correspond to a system operating at the left-hand edge of a normal recall-precision curve. Furthermore the queries, 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 user queries processed by SMART and Cranfield.

To summarize, no obvious evidence is in existence for distrusting the main results of the automatic indexing studies outlined earlier.

B) Comparison with Manual Indexing

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. [44,45] 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 determine under these circumstances how 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 abstract processing performed by SMART shows that the results obtained by the two systems are not statistically different. [53,57] To check these results, a comparison was

made later between the test results obtained by Lancaster for the manu ally-based Medlars system [30], and the SMART system. Specifically, for 18 of the Medlars queries used earlier by Lancaster, document abstracts were keypunched, and the retrieval process was repeated usin the automatic text searching methods incorporated into SMART. [63] The results indicate that for that subcollection a slightly higher average recall is obtained by SMART (0.695 compared with 0.643 for Medlars), whereas Medlars achieves a somewhat higher precision (0.611 for SMART and 0.625 for Medlars). In any case, the intuitive feeling that the conventional indexing would necessarily be superior is again not confirmed.

The results of the SMART-Medlars comparison might be interpret by saying that both the conventional and the automatic indexing procedures produce equally poor results (a recall and precision performance between 0.55 and 0.65 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 actually used. For the manual indexing process, Lancaster reports the following main sources of failure: [30]

index language problems (lack of specific terms or false coordination of terms);

search formulation (query formulation too exhaustive or too specific);

document indexing (document indexing insufficiently exhaustive, or too exhaustive, or omission of important terms);

lack of user-system interaction during 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 which 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 made first and, based on the preliminary output obtained, the search parameters are adjusted before attempting a second, more refined search. The adjustments made may then be different 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 [55,64,65] The first one is based on a selective print-out 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 requestor, 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 to be most helpful in improving the search results. Typically, the statistical term associations previously discussed 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 re-

phrasing the query in the user's hands. [64]

A second strategy consists in automatically modifying a search request by using the partial results from a previous search. Specifically, 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 purpose. Concepts from the documents termed relevant can then be added to the original search request if not present already, or their importance can be increased by a suitable adjustment of weights; contrariwise, terms from documents designated as irrelevant can be deleted or demoted. [55,64, 65,66] An illustration of such a relevance feedback process is shown in Fig. 3, where a query first retrieves a document identified as nonrelevant (Fig. 3 (a)). The query updating process which follows then shifts the query in such a way that a new search operation retrieves some relevant documents (Fig. 3 (b)). These in turn are used to generate two subqueries which are then successful in retrieving all relevant items (Fig. 3 (c)).

Considerable work has been done to optimize this type of feedback operation, and evaluation results indicate that the process produces considerable improvements in search effectiveness over the standard one-pass search process. A typical feedback evaluation graph showing averages for 200 documents and 42 queries in aerodynamics is shown in Fig. 4. Here an initial one-step search process using 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 query updating. The manual query updating is in turn compared with one iteration of the automatic relevance feedback

process. It is seen that the automatic query updating procedure is more effective than the manual one, and that an improvement of about twenty percent in precision is obtained through the feedback procedure. Moreover, this type of improvement in retrieval effectiveness is duplicated for all collections so far processed. [64,65,66] Under these circumstances, it appears safe to predict that future automatic information services will include interactive procedures for query or document modification during the search operation.

One more practical point dealing with document input and foreign language processing requires 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 at least abstract-length document excerpts should be available for analysis. Obviously, if all that material requires manual keypunching, the main benefits of the automatic analysis procedure may become 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 automatically readable document input products may well become generally available with each document before long.

Concerning the foreign language problem, the situation is less difficult than might appear. It is true that in certain subject areas, up to fifty 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 [30]). The English language analysis

methods will obviously not avail for these documents. However, it is also true that ninety percent of these documents are accounted for by only six or seven languages — most of them being in French, German, and Russian.

Some experiments were recently conducted with the SMART system using a collection of about 500 German documents in the field of library science. A multilingual thesaurus was prepared manually by translating the English version of an existing thesaurus into German. A thesaurus excerpt is shown in Fig. 5 from which it can be seen that the same concept class number represents both an English word class, as well as 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 query translation process. [67]

A sample German query processed through the German thesaurus is shown in Table 3. A comparison with Table 1 shows that a large number of "English" concepts are also present in the German analysis, thus accounting for the fact that the thesaurus translation is indeed successful. The foreign language problem does not then appear to present a major roadblock to an automatic document processing system.

7. Summary

A large number of automatic text analysis and indexing experi-

msnts 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 using syntactic analysis or hierarchical term expansion. Such a simple automatic indexing procedure is easily implemented on present-day computers, and no obvious reasons exist why manual document analysis methods should not be replaced by automatic ones.

While automatic document analyses appear therefore to be at least equivalent to 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 0.50 to 0.60 as at present. Further large-scale improvements are difficult to project. 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. [61,68] Such systems are still far removed from the ideal where both recall and precision are close to 1; furthermore, no obvious advances are likely to emerge which would produce such ideal systems.

Until such time as an ideal document handling system becomes available, presently known automatic document analysis and search methods provide suitable tools for document processing, and point the way to the fully automatic library of the future.

References

[1] Y. Bar Hillel, Theoretical Aspects of the Mechanization of Literature Searching, in Digitale Informationswandler, W. Hoffman, editor, F. Vieweg und Sohn, Braunschweig, 1962, p. 406-443 (citation p. 423).

[2] H. P. Luhn, A Statistical Approach to Mechanized Encoding and Searching of Literary Information, IBM Journal of Research and Development, Vol. 1, No. 4, January 1957.

[3] H. P. Luhn, Potentialities of Auto-Encoding of Scientific Literature, IBM Research Center, Report No. RC-101, Yorktown Heights, May 1959.

[4] J. O'Connor, Mechanized Indexing Methods and their Testing, Journal of the ACM, Vol. 11, No. 4, October 1964, p. 437-449.

[5] M. E. Stevens, Automatic Indexing: A State of the Art Report, U. S. Department of Commerce, NBS Monograph 91, Washington, March 1965.

[6] M. E. Stevens, V. E. Giuliano, and L. B. Heilprin, editors, Statistical Association Methods for Mechanized Documentation, Symposium Proceedings, U. S. Department of Commerce, NBS Monograph 269, December 1965.

[7] G. Salton, Automatic Information Organization and Retrieval, McGraw Hill Book Company, New York, 1968, 514 pages.

[8] G. Salton and M. E. Lesk, The SMART Automatic Document Retrieval System — An Illustration, Communications of the ACM, Vol. 8, No. 6, June 1965.

[9] G. Salton, Progress in Automatic Information Retrieval, IEEE Spectrum, Vol. 2, No. 8, August 1965.

[10] F. J. Damerau, An Experiment in Automatic Indexing, IBM Research Center, Report No. RC 894, Yorktown Heights, February 1963.

[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 View, G. Schecter, editor, Thompson Book Co., 1967.

[12] K. Sparck Jones and D. M. Jackson, The Use of Automatically Obtained Keyword Classifications for Information Retrieval, Final Report ML 211, Cambridge Language Research Unit, Cambridge, England, February 1969.

References (contd.)

[13] H. E. Stiles, Machine Retrieval using the Association Factor, in Machine Indexing: Progress and Problems, Third Institute on Information Storage and Retrieval, American University, Washington, February 1961.

[14] P. B. Baxendale, Machine-made Index for Technical Literature — An Experiment, IBM Journal for Research and Development, Vol. 4, No. 2, October 1958.

[15] D. J. Hillman and A. J. Kasarda, The Leader Retrieval System, Proceedings of the AFIPS Spring Joint Computer Conference, Thompson Book Co., May 1969.

[16] P. E. Jones, V. E. Giuliano, and R. M. Curtice, Papers on Automatic Language Processing — Development of String Indexing Techniques, Report ESD-TR-67-202, Vol. III, Arthur D. Little Inc., Cambridge, February 1967.

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

[18] M. J. Ruhl, Chemical Documents and their Titles: Human Concept Indexing vs. KWIC Machine Indexing, American Documentation, Vol. 15, No. 2, April 1964, p. 136-141.

[19] 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, p. 48-52.

[20] J. O'Connor, Correlation of Indexing Headings and Title Words in Three Medical Indexing Systems, American Documentation, Vol. 15, No. 2, April 1964, p. 96-104.

[21] H. Fangmeyer and G. Lustig, The Euratom Automatic Indexing Project, Proceedings of IFIP Congress 68, Edinburgh, August 1968.

[22] 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, p. 563-575.

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

[24] M. M. Kessler, Comparison of the Results of Bibliographic Coupling and Analytic Subject Indexing, American Documentation, Vol. 16, No. 3, July 1965, p. 223-233.

References (contd.)

[25] R. M. Needham, Applications of the Theory of Clumps, Mechanical Translation, Vol. 8, Nos. 3-4, June-October 1965.

[26] L. B. Doyle, Is Automatic Classification a Reasonable Application of Statistical Analysis of Text, Journal of the ACM, Vol. 12, No. 4, October 1965.

[27] H. Borko and M. D. Bernick, Automatic Document Classification, Journal of the ACM, Vol. 10, No. 2, April 1963.

[28] M. E. Maron, Automatic Indexing: An Experimental Inquiry, Journal of the ACM, Vol. 8, No. 3, July 1961, p. 404-417.

[29] J. O'Connor, Automatic Subject Recognition in Scientific Papers: An Empirical Study, Journal of the ACM, Vol. 12, No. 4, October 1965, p. 490-515.

[30] F. W. Lancaster, Evaluation of the Operating Efficiency of Medlars, Final Report, National Library of Medicine, January 1968.

[31] P. Atherton, D. W. King, and R. R. Freeman, Evaluation of the Retrieval of Nuclear Science Document References Using UDC as the Indexing Language for a Computer Based System, American Institute of Physics, Report AIP-UDC 8, May 1968.

[32] F. H. Barker and D. C. Veal, The Evaluation of a Current Awareness Service for Chemists, Chemical Society Research Unit in Information Dissemination and Retrieval, Report, August 1968.

[33] C. D. Gull, Alphabetic Subject Indexes and Uniterm Coordinate Indexes: An Experimental Comparison, in Studies in Coordinate Indexing, M. Taube and others, Documentation Inc., Washington, 1963.

[34] L. B. Heilprin and S. S. Crutchfield, Project Lawsearch: A Statistical Comparison of Coordinate and Conventional Legal Indexing, Proceedings of 1964 Annual Meeting of ADI, Spartan Books, Washington 1964, p. 215-234.

[35] M. R. Hyslop, Role Indicators and their Use in Information Searching, Relationship of ASM and EJC Systems, Proceedings of 1964 Annual Meeting of ADI, Spartan Books, Washington 1964.

[36] W. F. Johanningsmeier and F. W. Lancaster, Project SHARP Information Storage and Retrieval System: Evaluation of Indexing Procedures and Retrieval Effectiveness, Report NAVSHIPS 250-210-3, Bureau of Ships, Washington, June 1964.

[37] D. W. King, Evaluation of Coordinate Indexing Systems during File Development, Journal of Chem. Doc., Vol. 5, May 1965, p. 96-99.

References (contd.)

[38] D. B. McCarn and C. R. Stein, Intelligence Systems Evaluation, in Electronic Handling of Information: Testing and Evaluation, A. Kent and others, editors, Thompson Book Co., Washington, 1967, p. 110-122.

[39] E. Miller, D. Ballard, J. Kingston, and M. Taube, Conventional and Inverted Grouping of Codes for Chemical Data, Proceedings ICSI, Vol. 1, National Academy of Sciences — National Research Council, Washington 1959, pp. 671-685.

[40] B. A. Montague, Testing Comparison and Evaluation of Recall, Relevance, and Cost of Coordinate Indexing with Links and Roles, American Documentation, Vol. 16, No. 3, July 1965, p. 201-208.

[41] 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, Publication 1148, National Academy of Sciences, Washington, 1964.

[42] J. A. Schuller, Experience with Indexing and Retrieving in UDC and Uniterm, Aslib Proceedings, Vol. 12, November 1960, p. 372-389.

[43] J. Tague, Effectiveness of a Pilot Information Service for Educational Research Materials, Center for Documentation and communication Research, Western Reserve University, Cleveland, 1963, 56 pages.

[44] D. R. Swanson, Searching Natural Language Text by Computer, Science, Vol. 132, NO. 3434, October 21, 1960, p. 1099-1104.

[45] 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, p. 288-293.

[46] E. M. Fels, Evaluation of Performance of an Information Retrieval System by the Modified Mooers Plan, American Documentation, Vol. 14, No. 1, January 1963.

[47] B. Altmann, A Multiple Testing of the Natural Language Storage and Retrieval ABC Method: Preliminary Analysis of Test Results, American Documentation, Vol. 18, No. 1, January 1967.

[48] J. S. Melton, Automatic Processing of Metallurgical Abstracts for the Purpose of Information Retrieval, Final Report No. NSF-4, Center for Communication and Documentation Research, Case Western Reserve University, Cleveland, July 1967.

References (contd.)

[49] V. E. Giuliano and P. E. Jones, Study and Test of a Methodology for Laboratory Evaluation of Message Retrieval Systems, Report ESD-TR-66-405, Arthur D. Little Inc., Cambridge, August 1966.

[50] C. W. Cleverdon, The Evaluation of Systems used in Information Retrieval, Proceedings ICSI Conference, Vol. 1, National Academy of Sciences - National Research Council, Washington, 1959, p. 687-698.

[51] C. W. Cleverdon, Report on Testing and Analysis of an Investigation into the Comparative Efficiency of Indexing Systems, Cranfield, England, October 1962, 305 pages.

[52] G. Salton, The Evaluation of Computer-based Information Retrieval Systems, Proceedings of the FID 1965 Congress, Spartan Books, Washington, 1966.

[53] 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, 1966.

[54] C. W. Cleverdon, The Cranfield Tests on Index Language Devices, Aslib Proceedings, Vol. 19, No. 6, June 1967.

55 G. Salton, Search and Retrieval Experiments in Real-Time Information Retrieval, Proceedings IFIP Congress 68, North Holland Publishing Co., Amsterdam, 1969.

[56] G. Salton, The Evaluation of Automatic Retrieval Procedures — Selected Test Results using the SMART System, American Documentation, Vol. 16, No. 3, July 1965, p. 209-222.

[57] G. Salton and M. E. Lesk, Computer Evaluation of Indexing and Text Processing, Journal of the ACM, Vol. 15, No. 1, January 1968, p. 8-36.

[58] G. Salton, et. al., Information Storage and Retrieval, Reports No. ISR-11, ISR-12, ISR-13, ISR-14 to the National Science Foundation, Department of Computer Science, Cornell University, 1966-1968.

[59] C. A. Cuadra and R. V. Katter, Opening the Black Box of Relevance, Journal of Documentation, Vol. 23, No. 4, December 1967.

[60] A. M. Rees and D. G. Schultz, A Field Experimental Approach to the Study of Relevance Assessments in Relation to Document Searching, Final Report to the National Science Foundation, Center for Documentation and Communication Research, Case Western Reserve University, October 1967.

References (contd.)

[61] M. E. Lesk and G. Salton, Relevance Assessments and Retrieval System Evaluation, Information Storage and Retrieval, Vol. 4, No. 4, December 1968.

[62] T. Saracevic, The Effect of Question Analysis and Searching Strategy in Performance of Retrieval Systems: Selected Results from an Experimental Study, Comparative Systems Laboratory Report No. CSL:TR:15, Center for Documentation and Communication Research, Case Western Reserve University, Cleveland, May 1968.

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

[64] M. E. Lesk and G. Salton, Evaluation of Interactive Search and Retrieval Methods using Automatic Information Displays, Proceedings of the AFIPS Spring Joint Computer Conference, Thompson Book Co., Boston, May 1969.

[65] E. Ide, Relevance Feedback in an Automatic Document Retrieval System, Cornell University Master's Thesis, Report No. ISR-15 to the National Science Foundation, Department of Computer Science, Cornell University, January 1969.

[66] G. Salton, Search Strategy and the Optimization of Retrieval Effectiveness, in Mechanized Information Storage, Retrieval and Dissemination, K. Samuelson, editor, North Holland Publishing Co., Amsterdam, 1968, p. 73-107.

[67] G. Salton, Automatic Processing of Foreign Language Documents, in Information Storage and Retrieval, Report No. ISR-16 to the National Science Foundation, Department of Computer Science, Cornell University, 1969.

[68] C. W. Cleverdon, The Methodology of Evaluation of Operational Information Retrieval Systems based on the Test of Medlars, Report, Cranfield, England, June 1968.

Query Q 13 B (English)

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 Numbers	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

Thesaurus Analysis for English Query Q 13 B

Table 1

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

Order of Effectiveness of Three Types of Indexing Languages
(adapted from Cleverdon, et. al., [52] Fig. 8.1 T, p. 253)

Table 2

Query Q 13 B (German)

INWIEWEIT WERDEN COMPUTER-SYSTEME ZUR FORSCHUNG AUF DEM GEBIET DER SCHOENEN LITERATUR VERWENDET? HAT SICH MASCHINELLE SPRACHENANALYSE ALS HILFREICH ERWIESEN, UM Z. B. DIE VERMUTLICHE AUTORENSCHAFT BEI ANONYMEN WERKEN ZU BESTIMMEN ODER UM KONKORDANZEN ZUSAMMENZUSTELLEN?

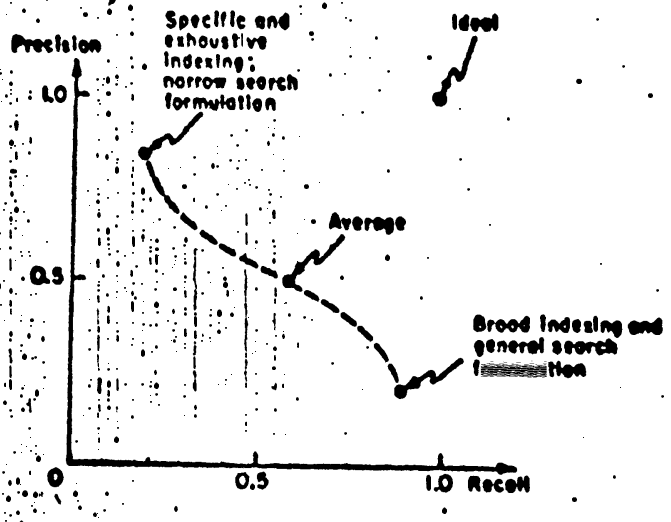
Concept Numbers	Weight	Sample Terms in Thesaurus Category
3 ✓	12	Computer, Datenverarbeitung
19 ✓	12	Automatisch, Kybernetik
21	4	Artikel, Presse, Zeitschrift
33 ✓	6	Analyse, Sprachenanalyse
45	4	Herausgabe, Publikation
64	4	Buch, Heft, Werk
65 ✓	12	Autor, Verfasser
68	12	Literatur
147 ✓	6	Linguistik, Sprache
207 ✓	12	Arbeitsgebiet, Fach
267 ✓	12	Konkordanz, KWIC
*		schoenen, hilfreich, vermutlich, anonymen, zusammenzustellen

✓ common concepts with English query

* query terms not found in thesaurus

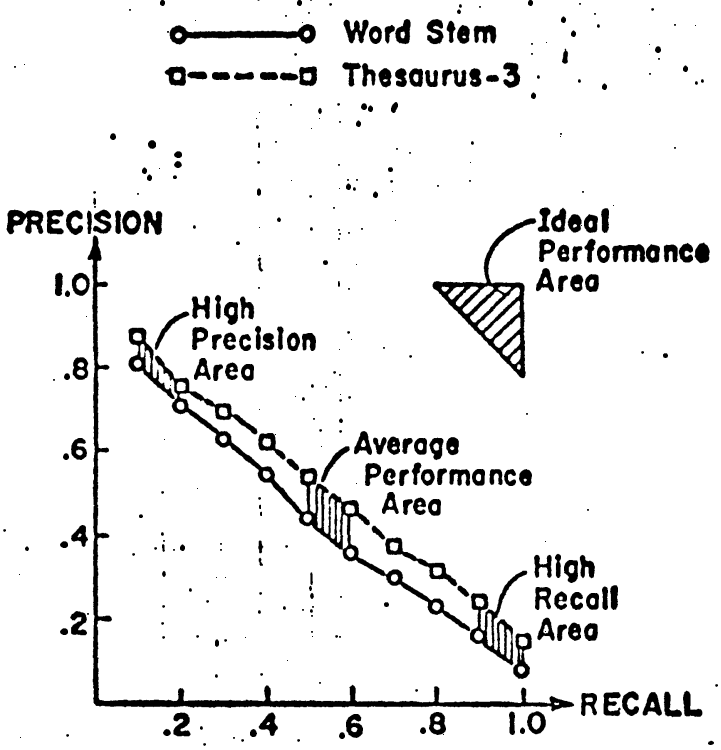
Thesaurus Analysis for German Query Q 13 B

Table 3



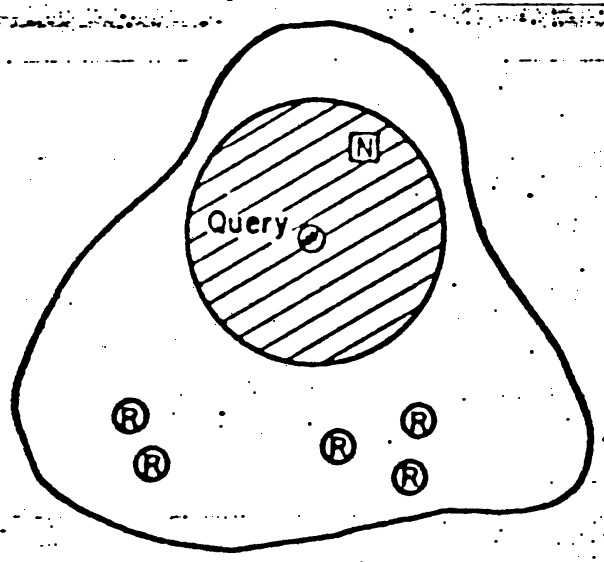
Performance Characteristics of Retrieval Systems

Fig. 1

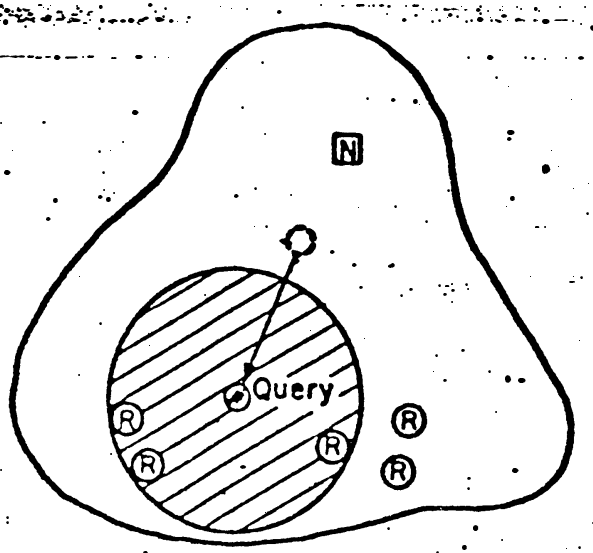


Word Stem – Thesaurus Comparison
(Averages for 780 IRE Documents)

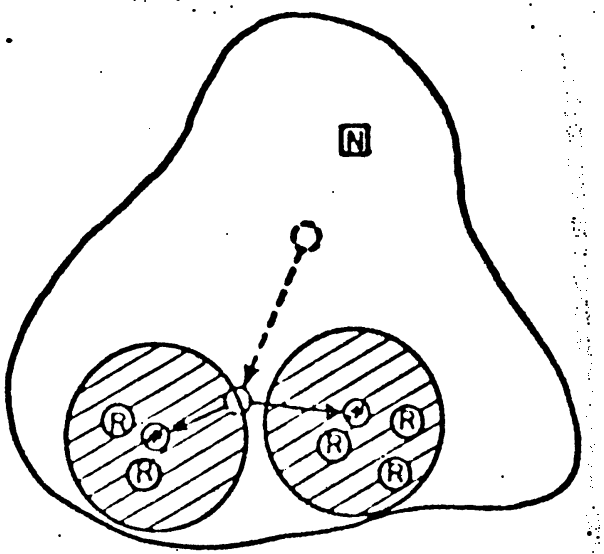
Fig. 2



a) Initial Query



b) First Alteration after Retrieving N



c) Second Alteration (Query Splitting)

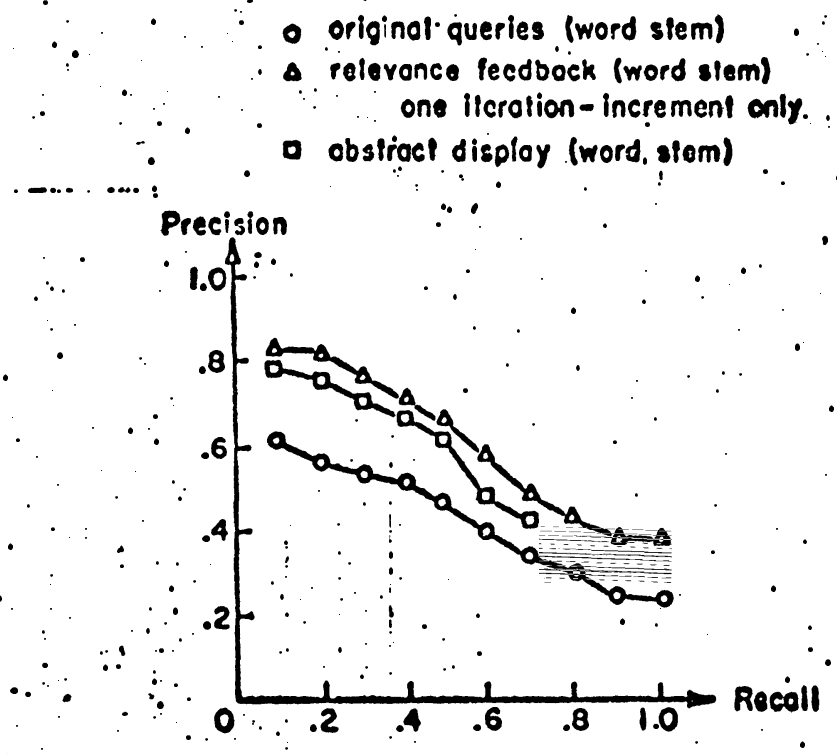
N Non-relevant Item

R Relevant Item

Query

Region of Retrieval

Relevance Feedback Illustration



Comparison of Abstract Display and Relevance Feedback

Fig. 4

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	

Multilingual English-German Thesaurus

Fig. 5