Another Look at Automatic Text Retrieval Systems*

Gerard Salton

TR 85-713
December 1985

Department of Computer Science
Cornell University
Ithaca, NY 14853

* This study was supported in part by the National Sciene Foundation under grant IST 83-16166.

Another Look at Automatic Text Retrieval Systems

Gerard Salton*

Abstract:

The characteristics of automatic text retrieval systems are briefly described, and the available experimental evidence comparing manual with automatic retrieval is reviewed. Several automatic text analysis and indexing models are then examined, and a basic automatic indexing process is proposed. There is no evidence that an intellectual content analysis performed by human subject experts produces better retrieval results than comparable automatic text processing systems.

1. Automatic Information Retrieval

An automatic text retrieval system is designed to search a file of natural language documents, and to retrieve certain stored items in response to queries submitted by a user population. Typically each stored item may be described by using for content identification certain words contained in the document texts, sometimes supplemented by additional related information. Queries are often formulated by using text words interrelated by the Boolean operators and, or, and not. The retrieval system is then designed to retrieve all stored texts identified by an appropriate combination of query words.

For example, if the user were interested in obtaining information about the design of small computers, the query could be formulated as [(minicomputers or microcomputers or hand-held calculators) and (design or construction or

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

architecture)]. The retrieval system would then extract from the file items containing the identifiers "design" and "minicomputers", or the identifiers "construction" and "microcomputers". [1,2]

It is customary to evaluate the effectiveness of a retrieval system by using a pair of measures, known as recall and precision, respectively. Recall is the proportion of relevant material actually retrieved from the file, and precision is the proportion of the retrieved material which is found to be relevant to the user's information needs. In principle, a search should achieve high recall by retrieving most everything that is relevant, while at the same time maintaining a high precision by rejecting a large proportion of the extraneous items. In that case, both the recall and the precision values of the search are close to 1 (or 100 percent). In practice, it is known that recall and precision tend to vary inversely, and that it is difficult to retrieve everything that is wanted, while also rejecting everything that is unwanted.

In particular, when very specific query formulations are used few non-relevant items tend to be obtained, but also relatively few relevant ones. That is, a specific query formulation produces a high precision, and hence low recall, performance. As the query formulation is broadened, more relevant items are retrieved, thus improving the recall, but also more nonrelevant ones, depressing the precision. In that case, one may obtain high recall but also low precision. A compromise is often reached in practice by using a query formulation which is neither too narrow nor too broad in the hope of obtaining both reasonable recall and reasonable precision values. When a choice must be made between recall and precision, most users choose precision oriented searches because only relatively few items may then be retrieved,

thus limiting the user effort needed in examining the retrieved materials. A penalty in user time and effort is attached to a high recall search, because the higher the desired recall, the more retrieved items must normally be examined.

In automatic retrieval systems, query formulations and document representations can be altered in attempting to reach desired recall and precision levels. In particular, the use of recall-enhancing devices will broaden the document and query identifiers in the hope of achieving a higher recall performance. Analogously, the precision-enhancing devices are designed to render the item identifications more specific in the expectation of obtaining better precision. A typical list of recall and precision enhancing devices appears in Table 1.

A relatively simple method for reaching a higher recall consists in using truncated terms, or word stems, instead of the original complete terms, for query or document identification. A form such as "analy" covers notions such as "analyst", "analysis", "analyzer", etc., and has a broader scope than any of the complete words. Other recall enhancing devices include the use of terms that are synonymous or related to the original ones, and the use of broader, more general terms than the original ones. Such terms can be obtained from previously available thesauruses and term hierarchies, or they can be suggested by users during the search operations.

The use of term weights may enhance the search precision, because the weights distinguish the better, or more important, terms from the less important ones. Such a discrimination may help in ranking the output in decreasing order of presumed importance. Other precision oriented devices consist in using term phrases instead of single terms -- for example, "computer

programmer" instead of "computer" -- and supplying narrower terms obtained from a term hierarchy. Useful term phrases could again be contained in a previously available dictionary, or they could be formed from sets of single terms that cooccur in the documents of a collection.

Most automatic text retrieval systems make provision for the use of truncated terms, and the addition of related terms. Some automatically generated term weights may also be used to distinguish the items containing the more highly weighted terms from others containing terms of lower weight.

2. The Blair and Maron Retrieval Test

In a recent paper by Blair and Maron (BM), the well-known automatic text retrieval system STAIRS was examined, using a collection of 40,000 full text documents, equivalent to some 350,000 pages of text, and 40 different user queries. [3] The STAIRS system normally uses words extracted from document texts for content identification. The text words may be broadened by using truncation, and each word may be supplemented by lists of synonyms supplies by the users. When synonyms are specified, a search based on a particular term automatically extends to the whole synonym list. The STAIRS system also includes a ranking feature capable of retrieving documents in decreasing order according to the sum of the weights of matching query terms contained in each document. Term weights are computed by using the occurrence frequencies of the terms in the retrieved document subset. [4]

While some features of the STAIRS system are not as attractive as they could be -- for example, a more reasonable term weighting system might produce better retrieval performance -- STAIRS is certainly a state-of-the-art full text retrieval system, and its operations are typical of what is obtainable

with existing, operational, automatic text search systems. In the retrieval test, Blair and Maron obtained an average precision value of about 75 percent (0.75) and an average recall value of 20 percent (0.20). That is, for each of the 40 test searches, 3 out of 4 retrieved documents were in fact pertinent to the user queries, and approximately one fifth of the total number of relevant items present in the collection was retrieved.

As will be seen, such a performance may be typical of what is achievable in existing, operational retrieval environments. Furthermore, for the many retrieval system users who cannot spend many hours examining large masses of retrieved materials, the STAIRS performance reported by Blair and Maron should represent a high order of retrieval effectiveness. Indeed, the searchers were able to extract from a large collection of 40,000 documents a substantial number of useful items, and since only 1 of 4 retrieved items proved to be extraneous, the time wasted in considering useless items must have been comparatively small.

Unfortunately, in the case of the BM test, the searchers were lawyers, and the materials being searched were legal documents. In view of the Anglo-American legal system based on common law and judicial precedence, many lawyers are of necessity high-recall users. That is, in order to know how a particular legal case needs to be approached, it is often important to examine all possible previous cases that may be similar in some sense to the current case, because a legal argument made in a related previous case could be applicable to the current one. The high precision search output obtained by the BM test searches, which rejected most nonrelevant materials but obtained only about twenty percent of the potentially useful items, might be fitting for research workers, or for university professors and students. However, for the

legal personnel that actually conducted the searches, a better recall performance was thought to be mandatory even at the cost of a decreased search precision.

Blair and Maron derive three main conclusions from their retrieval test: [3]

a) When high recall is desired, the searchers of large collections cannot simply broaden the search requests as would be done for small collections used experimentally in the laboratory. When broader search formulations are used, the search precision might suffer intolerably, and the user could be swamped with masses of useless retrieved materials. For this reason, some earlier test results which showed the superiority of text based retrieval over manual systems for small collections are not applicable in the real world of large collections:

> "Retrieval strategies that work well on small systems do not necessarily work well on larger systems (primarily because of output overload)"

b) When high recall is desired, the use of manually, or intellectually, indexed collections is preferred over systems based on the full document texts:

> "...the full text system means the additional cost of inputting and verifying 20 times the amount of information that a manually indexed system would deal with. This difference alone would more than compensate for the added time needed for manual indexing and

vocabulary construction."

c) The full text systems, and particularly STAIRS, are not user friendly since even the trained searchers used in the experiment obtained inadequate performance; untrained users would no doubt do worse:

> "STAIRS is sold under the premise that it is easy to use and requires no sophisticated training on the part of the user ..."

Given the impressive precision performance of the STAIRS system in the BM test environment, the BM report ends with a surprising paraphrase of Samuel Johnson: "full text searching is one of those things that ... is never done well, and one is surprised to see it done at all." ([3], p. 298).

In reacting to the cited conclusions of the BM study, one must consider that no comparison was made in the study between full text retrieval systems and manually indexed systems, nor was a comparison made of the retrieval performance of large versus small document collections. The cited conclusions are thus not based on any data submitted by the authors, nor are they supported by other evidence available elsewhere. In fact, as will be seen, there exists substantial evidence to show that these conclusions are more sentiment than fact:

a) The evidence for several retrieval evaluations conducted with very large document collections does not support the notion of output overload, although of course high recall implies more retrieved items, and hence more work in analyzing the output than low recall searches.

b) Comparisons between manual and automatic indexing systems have been performed for large document collections and these tests indicate that the automatic text based systems are at least competitive, or even superior to the systems based on intellectual indexing. These results are ignored in the BM report.

c) Automatic indexing theories exist which provide index terms that are not simply words extracted from document texts. Indeed the automatic indexing results by Swanson and Salton [5,6] that are cited in the BM study were not based on the use of full document texts, but on the analysis of document abstracts; the automatic indexing systems can outperform manual indexing systems without requiring an analysis of full document texts.

In the remainder of this note, performance results are given for other automatic retrieval experiments using large document collections, and conclusions are drawn regarding suitable automatic indexing methodologies.

3. Experiments with Large Retrieval Systems

A) The Medlars Evaluation

In the late nineteen sixties, Lancaster conducted an in-house study of the Medlars demand search service which is operated by the National Library of Medicine in Bethesda, Maryland, for literature in the biomedical field. [7] Medlars uses manual, or intellectual indexing performed by subject experts, using a controlled indexing language based on the Mesh (Medical Subject Headings) thesaurus. Following a manual indexing operation of the stored documents and a manual query formulation, the file search and retrieval operations

are performed automatically.

The in-house evaluation of Medlars involved a data base of over 700,000 documents in biomedicine, and a set of about 300 test queries. The search results varied widely, in the sense that some queries performed perfectly (recall = 1.00 and precision = 1.00), while others retrieved no relevant items at all (recall = 0 and precision = 0). For the 300 queries, the average recall performance was 0.58 while the average precision was 0.50. For documents of "major value", the average recall increased to 0.65. In presenting the results, Lancaster remarks that the actual performance value obtained for a query can be made to vary by submitting more or less specific query formulations. The average performance for a query set can then be made to slide along a monotonically decreasing curve starting at the high precision-low recall end of the performance spectrum, and going down to the high recall-low precision end as the query formulations are broadened. The resulting curve representing the performance of the Medlars search system is shown in Fig. 1. A second, lower curve also included in Fig. 1, represents the 75th percentile curve, giving the performance points exceeded for 75 percent of the test queries.

Three particular performance points for Medlars are analyzed in more detail in Table 2. It may be noted that for the high precision searches, the precision performance was about 0.80 but the recall reached only 0.19. For these searches about 50 items were retrieved (out of some 700,000), of which about 40 were relevant. At the average performance point of 0.58 recall and 0.50 precision, the retrieved set increases to 175 documents of which about 60 percent were relevant on average. For high recall searches, the recall reached nearly 90 percent (0.89) but the precision dropped to 0.20. To obtain

that recall performance it was necessary to retrieve between 500 and 600 items out of 700,000 of which about 130 on average were relevant to the query. The feared "output overload" predicted by Blair and Maron does thus not occur for the Medlars search service. This is not likely to be due to the intellectual indexing, but rather to the heterogeneity of the collections which cover all of biomedicine thus simplifying the rejection of useless materials for any one search.

The set of 500 items retrieved on average for the high recall searches represents only seven one hundredth of a percent of the collection; nevertheless such a high recall output entails substantial work for the users, and only specially motivated users (such as lawyers) will want to submit such broad query formulations to a search service. Lancaster makes the following comments in this connection:

> "We can choose to operate Medlars, as it presently exists, at any performance point on or near the recall-precision plot (of Fig. 1) ... Intuitively one feels that Medlars should be operating at a higher average recall ratio (than 0.58) and should sacrifice some precision in order to attain improved recall. However Medlars is now retrieving an average of 175 citations per search in operating at recall 0.58 and precision 0.50. To operate at an average recall of 85 to 90 percent and an average precision of 20 to 25 percent implies that Medlars would need to retrieve an average of 500 to 600 citations per search. Are requestors willing to scan this many citations to obtain a higher level of recall?" [7]

The performance point obtained in the BM study for the STAIRS text search

system (0.75 precision, 0.20 recall) is superimposed on the Medlars performance curve of Fig. 1. It may be noted that the STAIRS performance falls well within the range of the high precision searches of Medlars, even though no controlled indexing language is used by STAIRS, and no intellectual indexing. The query broadening, recall-enhancing devices listed in Table 1 are of course available in an automatic environment such as STAIRS just as they are in the controlled language environment of Medlars.

The recall and precision failure analysis undertaken by Lancaster for the Medlars searches shows that plenty of problems arise also in manual indexing environments. A summary of the failure analysis for 797 recall failures (failures to retrieve relevant items) and 3038 precision failures (failures to reject nonrelevant items) appears in Table 3. Some of the failures included in Table 3 have multiple causes accounting for totals that may exceed 100 percent.

The data of Table 3 show that a substantial proportion of the search failures are due to the human indexing and to the controlled indexing language used in the Medlars environment. Some of these failures might be avoidable in an automatic indexing situation while others would not. Poor search formulations and inadequate user-system interaction may occur with any retrieval system whether manual or automatic. However, the conventional, manual retrieval system may well be especially vulnerable, as shown by the following comments by Cleverdon: [8]

> "If two people or groups of people construct a thesaurus in a given subject area, only 60 percent of the index terms may be common to both thesauri;

if two experienced indexers index a given document using a given thesaurus, only 30 percent of the index terms may be common to the two sets of terms;

if two search intermediaries search the same question on the same database on the same host, only 40 percent of the output may be common to both searches;

if two scientists or engineers are asked to judge the relevance of a given set of documents to a given question, the area of agreement may not exceed 60 percent."

"There will be many exceptions to the above generalized statements, but all are supported by results from experimental or operational tests, and under such circumstances it is surprising that (manual) retrieval systems can operate at performance levels (as high as) 60 percent recall and 50 percent precision."

Cleverdon offers a solution for this unfortunate situation which he spells out in the following terms: [8]

"The problems caused by the use of a controlled language thesaurus and variations in (manual) indexing can be overcome by eliminating these two activities and using, as the input, an extract such as the title and abstract in natural (or free-text) language. Basically, a controlled language represents a reduction in the totality of the potentially available terms in the given subject area ... (due to) compounding of real synonyms or spelling variations ... (or to) subsuming of one or more specific terms by a general term ....."

"Such combining of search terms as may, in a given search, be considered necessary is better done at the search stage than at the input. This appears to be one of the reasons why, in every test which has compared the performance of searching on controlled language index terms as against searching on abstracts in natural language, the results have been in favor of natural language."

Some of the test results comparing natural language text searching with manual indexing systems are reviewed in the next subsection.

B) Comparison of Manual and Automatic Indexing

In the middle nineteen seventies a comparison between automatic and manual indexing was made using a NASA database consisting of documents from the Scientific and Technical Aerospace Reports (STAR) and the International Aerospace Abstracts (IAA). The test was based on a collection of 44,000 document titles and abstracts processed against 40 search requests. The following indexing systems were compared:

a) a natural language text search system consisting of a machine search of document titles and abstracts (not of full text);

b) a natural language text search system supplemented by a thesaurus of "associated concepts" prepared from the source documents;

c) a controlled language indexing performed by human subject experts;

d) and finally, the controlled indexing supplemented by natural language terms extracted from the documents.

The search results for the NASA test are summarized in Table 4. It may be seen that the natural language abstract search produces the best average recall for the 40 test queries (0.78) and a high order of precision (0.63). The controlled language manual indexing produced a better precision value than the automatic abstract search (0.74) but a substantially worse recall (0.56). When these results are considered, it is certainly not possible to reach the conclusion that searches of natural language abstracts are inferior to the controlled language indexing. Indeed if the NASA search population had consisted of legal personnel with a recall orientation similar to the searchers involved in the Blair and Maron test, the NASA searchers would certainly have preferred the output produced by the automatic search system with its recall advantages of over 20 percent over the manual system. Cleverdon, who was in charge of the NASA system test reaches the following conclusion: [9]

> "it appears impossible to reach any other conclusion than that, within the parameters of this test, natural language searching on titles and abstracts proved at least equal to, and probably superior to, searching on controlled language terms; it also seems that a significant factor in this (result) was the increased level of indexing exhaustivity (provided by the natural language text search system)."

The performance points for the NASA search system evaluation are shown in Fig. 2. The curve corresponding to the controlled term performance for the in-house Medlars text shown earlier in Fig. 1 is also superimposed on the output of Fig. 2. A comparison of the NASA output with the STAIRS performance involving collections of comparable size shows that the NASA searches are substantially more effective. Collection size does not seem to play a substantial role in the search performance.

Query type and homogeneity in the subject matter of the collection are likely to more important.

Many additional comparisons between automatic indexing and controlled term indexing systems appear in the literature. For example, a small sample collection of 450 documents and 29 search requests was used to compare the performance of the intellectual controlled-term Medlars system with an automatic indexing system based on abstract searching supplemented by the use of a thesaurus of related terms. [10] The two competing systems produced almost identical results for the test collection (0.31 recall and 0.61 precision for the controlled term indexing versus 0.32 recall and 0.61 precision for the natural language terms plus thesaurus).

In the well-known Aslib-Cranfield study an attempt was made to evaluate the performance of natural language "single term" indexing based on abstract searching, and supplemented by many types of recall and precision enhancing devices. The automatically derived single term languages were then compared with various kinds of intellectual controlled term indexing systems. [11] A sample collection of 1400 documents in aeronautics was used for test purposes in the Aslib Cranfield study, together with 221 test queries. As the two typical performance curves included in Fig. 2 show ([11], p. 127 and p. 164), the recall-precision performance for the Cranfield collection was relatively poor, compared with other previously mentioned results obtained for much larger test collections. However, in practically every case, the Aslib Cranfield tests showed that the single term natural language indexing provided somewhat better search results than the comparable controlled term indexing. This is the case also for the two illustrated Cranfield searches included in Fig. 2.

It was mentioned already that an automatic text search system does not

need to restrict itself to the use of single words extracted from document texts. Complete automatic indexing packages can be used to construct automatic document representations used for search and retrieval. A brief summary of automatic indexing theory and practice appears in the next section.

3. Automatic Indexing Theories

The effectiveness of an indexing system designed to produce useful content representations for written texts depends on two main characteristics: the exhaustivity of the indexing, that is, the degree to which all aspects of the document content are recognized and represented in the indexed document representations, and the specificity of the individual index terms used to represent document content, that is, the level of detail at which a given content- or index term is represented. A high degree of exhaustivity of the indexing may improve the recall performance of a search by permitting the identification of relevant materials that would remain unrecognized when the indexing exhaustivity is lower. On the other hand, a high degree of specificity in the index terms is likely to favor search precision.

In principle, the choice of an indexing system useful for the content representation of natural language texts should be based on linguistic considerations, including especially also semantic components. Linguistic analysis methods are, however, difficult to apply efficiently to large text samples, and for this reason most existing indexing theories are based on statistical or probabilistic methodologies. On the simplest level, both indexing exhaustivity and index term specificity may be characterized by using the occurrence statistics of the terms in the documents of a collection. In particular, the exhaustivity of the indexing is characterized to some extent by

the number of index terms assigned to a given document, whereas term specificity may be related inversely to the number of documents in a collection to which a term is assigned. [12] Thus terms that are assigned rarely may be assumed to be more specific than other more frequently assigned entities.

In judging the value of a term for purposes of content representation two different statistical criteria may be of interest: On the one hand, a term that appears often in the text of an individual document, or of a document excerpt, may be assumed to carry more importance for the content representation of that document than a more rarely occurring term. Thus a document containing the term "pear" many times is likely to deal with the notion of pears. On the other hand, if that same term also occurs in many other documents of the collection -- that is, if all other documents also deal with pears -- then the term "pear" may not be as valuable as other terms that occur more rarely in the remaining documents of the collection. This suggests that the goodness of a given term assigned to a given document be measured as a combination of its frequency of occurrence inside that document (the term frequency, tf), and an inverse function of the number of documents in the collection to which it is assigned (the inverse document frequency, idf). The idf factor could be computed, for example, as 1 divided by the document frequency. A possible term weighting function for term i in document j may then be

$$w_{ij} = tf_{ij} \cdot idf_i \quad . \tag{1}$$

The first factor in the product of expression (1) is document dependent whereas the latter factor is collection dependent. [13] Using this term importance strategy, the best terms assigned to the documents will be those occurring frequently inside particular documents but rarely on the outside. Such terms may in fact be used to distinguish the documents of a collection

from each other. Both factors of the product are easy to utilize because the inverse document frequency of a term can be obtained in advance from a collection analysis, whereas the term frequencies can be computed from the individual documents, as needed.

The well-known probabilistic retrieval model is based on the premise that the most valuable documents for retrieval purposes are those whose probability of relevance to a query is largest. [14] The relevance properties of the documents can be estimated by using the relevance properties of the individual terms included in the documents, and those in turn can be obtained by looking at the frequencies of the terms in the sets of relevant and nonrelevant documents of a collection. When the terms are assumed to occur independently of each other in the documents of a collection -- an assumption which is not always realistic in practice -- the probabilistic retrieval model leads to a term weighting function which can be shown to be optimal under the assumptions of the model, known as the term relevance weight tr_i for term i: [15]

$$tr_i = \log \frac{p_i(1-q_i)}{q_i(1-p_i)} \quad . \tag{2}$$

The parameters p_i and q_i represent the probabilities of occurrence of term i in the relevant and nonrelevant documents of a collection, respectively.

The relevance probabilities of the terms might be estimated by using an iterative search process which allows the user to supply relevance judgments for certain documents retrieved in earlier searches of the collection. Alternatively, some simplifying assumptions may be made, based on the premise that the set of relevant documents for a particular query is likely to be very small compared to the collection size, and that the probability of relevance p_i for the various terms i can then be assumed to be a constant. [16] In

that case, the term relevance weight tr_i for term i of expression (2) reduces to $\log((1-q_i)/q_i)$. Since the set of nonrelevant documents in a collection is approximately equal to the collection size N, the probability of nonrelevance q_i of term i is approximately equal to the probability of occurrence of term i in the whole collection, or n_i/N, where n_i represents the number of documents in the collection with term i. Under those simplifying assumptions, the term relevance formula of expression (2) is then reduced to the form

$$tr_i = \log\frac{N-n_i}{n_i} + \text{constants} \quad . \tag{3}$$

which can be computed without knowing anything about the relevance of particular documents to particular queries.

The formula of expression (3) represents an inverse document frequency (idf) factor, because the higher the document frequency n_i of a term, the lower will be the relevance weight tr_i. The probabilistic retrieval model thus provides some justification for the use of the inverse document frequency factor in expression (1), because under appropriate assumptions the idf_i factor is approximately equal to the optimal term weight tr_i.

A somewhat different, but related way of approaching the document indexing task in a retrieval environment may be based on the term discrimination model. [13] In this model, the assumption is made that the most useful terms for the content identification of natural language texts are those best capable of distinguishing the documents of a collection from each other. This suggests that the value of a term should be measured by calculating the decrease in the "density" of the document collection which results when a given term is assigned to the documents of a collection. The density of the document space reflects the degree to which the document representations

resemble each other. It could be measured by computing, for example, the sum of the pairwise document similarities for all pairs of documents in the collection. Using this strategy, the density of the documents will be high when the documents resemble each other a great deal, that is, when they are indexed by many of the same terms.

Under the term discrimination approach, the broad high frequency terms become the least desirable content identifiers, because they will be assigned to many documents of a collection, thereby enhancing the mutual similarity of the corresponding documents. When a broad high-frequency term is assigned, the average similarity between documents, and hence the document space density, increases. If the discrimination value of a term is measured as the collection density before the given term assignment minus the density after term assignment, it is clear that high frequency terms are characterized by a negative term discrimination value. In the term discrimination model, the very rare, low-frequency terms that are preferred by the inverse document frequency factor are also not very desirable for content identification, because such terms are assigned to very few documents of a collection, so that the space density hardly changes when those such terms are introduced. The very rare terms thus receives a discrimination value close to zero.

The best content identifiers will be those occurring neither too rarely nor too frequently. Such terms may be assigned to as many as one tenth of the items in the collection, and they serve to distinguish the items to which they are assigned from the remainder. A graphic representation of the variations in term discrimination value is shown in Fig. 3 as a function of the document frequency of the terms. As the number of documents to which a term is assigned increases from zero, the term discrimination value first increases

from zero and becomes positive; eventually, as the document frequencies become still larger, the term discrimination values decrease rapidly and become negative for high frequency terms.

The term discrimination model confirms the notion that a correct degree of specificity exists for terms used as content identifiers of texts, and that terms which do not exhibit the appropriate specificity should be broadened when too specific, or narrowed when too broad. [12] The recall and precision-enhancing devices included in Table 1 may conveniently be used for that purpose. A principal method of term broadening consists in using a thesaurus, or other vocabulary grouping device, to supply synonyms and related terms of various kinds, whereas a simple method for term narrowing consists in introducing term phrases to replace certain broad single terms. The term discrimination model thus assigns a specific role to a thesaurus as a grouping device for related narrow terms. Term phrases are used as always to render broad concepts more specific. The thesaurus and phrase transformation methods produce shifts of the terms toward the center of the frequency spectrum where the content identifiers with the best specificity are located. The left-to-right theasurus and right-to-left phrase transformations are indicated on the graph of Fig. 3.

4. A Blueprint for Automatic Indexing

The preceding examination of automatic indexing theories leads to the design of effective automatic text-based retrieval systems that are competitive with conventional manual operations and can be operated without human subject experts for document indexing and search formulation. The following basic process suggests itself for this purpose: [17]

1) Identify the individual words occurring in the documents of a collection, or in chosen document excerpts such as titles and abstracts.

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

3) Use a suffix stripping routine to reduce the remaining words to word stem form; such a recall-enhancing transformation broadens the scope of the terms and can be performed automatically using a limited number of basic rules. [18]

4) For each remaining word stem i occurring in document j compute a term weighting factor as the product of the term frequency of term i in document j multiplied by the inverse document frequency of term i in the collection; the available evaluation results indicate that term weighting improves retrieval effectiveness by distinguishing the important content terms from the less important ones. [19]

5) Represent each document by the chosen set of weighted word stems.

The retrieval evaluation results available for this type of simple indexing for both large and small document collections indicate that this "single-term" indexing method is competitive with and often superior to conventional intellectual indexing systems. [9-11] The STAIRS system used in the test by Blair and Maron makes provisions for all steps of the basic system with the exception of the term weighting. In STAIRS, term weights may be assigned after retrieval of the documents based on term

occurrence characteristics in the retrieved document subset only. The STAIRS weighting is then used to generate a ranked list of the retrieved documents. The use of ranked document output facilitates user-system interaction by letting the user see the more important documents first. Information from the documents retrieved early in a search can then be used to generate improved query formulations to be processed in subsequent searches.

Ideally, the term weights should, however, be generated before the query and document representations are compared during the search, and the term weights should be computed by using complete collection statistics instead of information from a particular subset (the retrieved items) that is not representative of the whole collection. For example, terms exhibiting high occurrence frequencies in the retrieved subset can certainly not be labelled as good or bad, unless something is known also about their occurrence frequencies in the remainder of the collection.

The basic indexing process can be improved by adding the following refinements:

1) Generate weighted word stems attached to the documents as described earlier.

2) Use a thesaurus to replace terms with low document frequencies (and near zero discrimination values) by the corresponding thesaurus class identifications.

3) Use a phrase formation process to generate term phrases incorporating terms with high document frequencies (and negative discrimination values) based on term cooccurrences in the

document excerpts.

4) Compute a combined term weight for assigned thesaurus classes and term phrases, and represent each document by the corresponding sets of weighted single terms, term phrases and thesaurus classes.

In the STAIRS system, the thesaurus is generated "on the fly" by letting the user suggest terms that are snyonymous or related to particular index terms. These related terms are then used automatically to "expand" the set of original terms. A previously available thesaurus which groups low frequency terms into classes of related terms could be used for the same purpose.

A natural language query formulation can be converted into sets of weighted terms in the same way as a document text. Composite query-document similarity coefficients can then be computed, reflecting the similarities between corresponding term representations. When query-document similarity measurements are available, the documents can be ranked for output purposes in decreasing order of the query-document similarity. Furthermore improved query formulations can be generated by using information obtained from the texts of previously retrieved documents. [17]

When the search requests are submitted in Boolean form, as they are in many operational retrieval environments, weighted terms can also be incorporated, and an approximate, fuzzy match between the weighted term sets representing the documents, and the weighted Boolean query statements can be used to produce a ranked output in decreasing order of

similarity to the queries as described earlier for queries consisting of term sets without Boolean operators. [20-21]

No support is found in the literature for the claim that text-based retrieval systems are inferior to conventional systems based on intellectual human input. Indeed, all the available evidence obtained with large and small collections indicates that properly designed text based systems are preferable to manually indexed systems. Furthermore, as Swanson has pointed out over 25 years ago "..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 systems." [5]

References

[1] F.W. Lancaster, Information Retrieval Systems: Characteristics, Testing, and Evaluation, Second Edition, J. Wiley and Sons, New York, 1979.

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

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

[4] IBM World Trade Corporation, Storage and Information Retrieval System (STAIRS)-- General Information Manual, Second Edition, IBM Germany, Stuttgart, Germany, April 1972.

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

[6] G. Salton, Automatic Text Analysis, Science, 168:3929, April 1970, 335-343.

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

[8] C.W. Cleverdon, Optimizing Convenient On-Line Access to Bibliographic Databases, Information Service and Use, 4, 1984, 37-47.

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

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

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

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

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

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

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

[16] W.B. Croft and D.J. Harper, Using Probabilistic Models of Document Retrieval without Relevance Information, Journal of Documentation, 35:4, December 1979, 285-295.

[17] G. Salton, A Blueprint for Automatic Indexing, ACM SIGIR Forum, 16:2, Fall 1981, 22-38.

[18] J.B. Lovins, Development of a Stemming Algorithm, Mechanical Translation and Computational Linguistics, 11:1-2, March and June 1968, 11-31.

[19] G. Salton and M.E. Lesk, Computer Evaluation of Indexing and Text Processing, Journal of the ACM, 15:1, January 1968, 8-36.

[20] G. Salton, E.A. Fox and H. Wu, Extended Boolean Information Retrieval, Communications of the ACM, 26:11, November 1983, 1022-1030.

[21] G. Salton, A Blueprint for Automatic Boolean Query Processing, ACM SIGIR Forum, 17:2, Fall 1982, 6-25.

Recall Enchancing Devices (term broadening)	Precision Enhancing Devices (term narrowing)
term truncation (suffix removal)	term weighting
addition of synonyms	addition ot term phrases
addition of related terms	use of term cooccurrences in documents or sentences
addition of broader terms (using term hierarchy)	addition of narrower terms (using term hierarchy)

Typical Recall and Precision Enhancing Devices

Table 1

Performance Points	Recall	Precision	Number of Retrieved Items	Number of Relevant Retrieved
high precision searches	0.19	0.80	40-50	30-40
medium performance	0.58	0.50	175	85
high recall searches	0.20	0.89	500-600	135

Medlars Performance Points

Table 2

Source of Failure	797 Recall Failures	3038 Precision Failures
Indexing Language (lack of appropriate term, false coordination)	10.2%	36.0%
Search Formulation (too specific, or too exhaustive)	35.0%	32.4%
Document Indexing (too specific, or too exhaustive)	37.4%	12.9%
Inadequate User-System Interaction	25.0%	16.6%

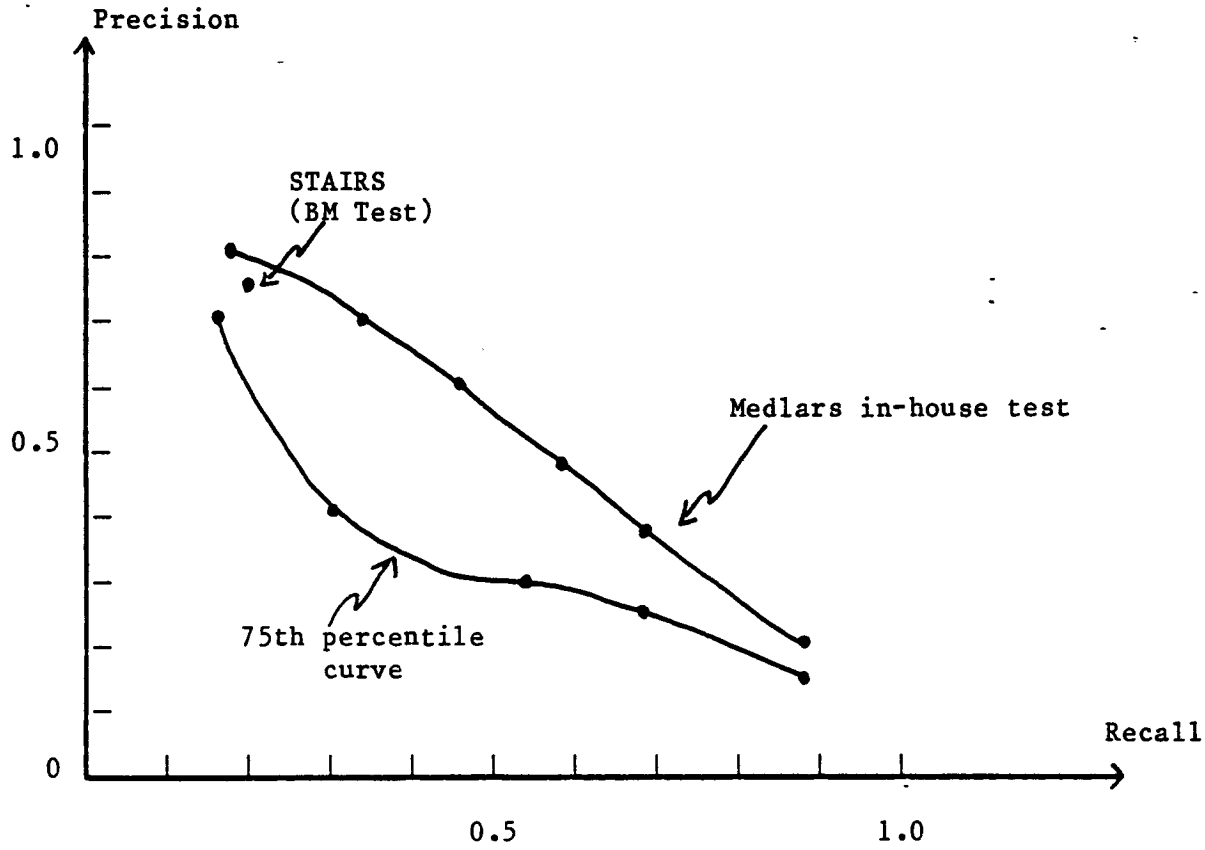
Typical Failures of Medlars Searches
(adapted from [7])

Table 3

Indexing Method	Recall	Precision
Natural language indexing (text search of titles and abstracts)	0.78	0.63
Natural language supplemented by associated concepts	0.73	0.52
Controlled language manual indexing	0.56	0.74
Controlled language supplemented by natural language terms	0.71	0.45

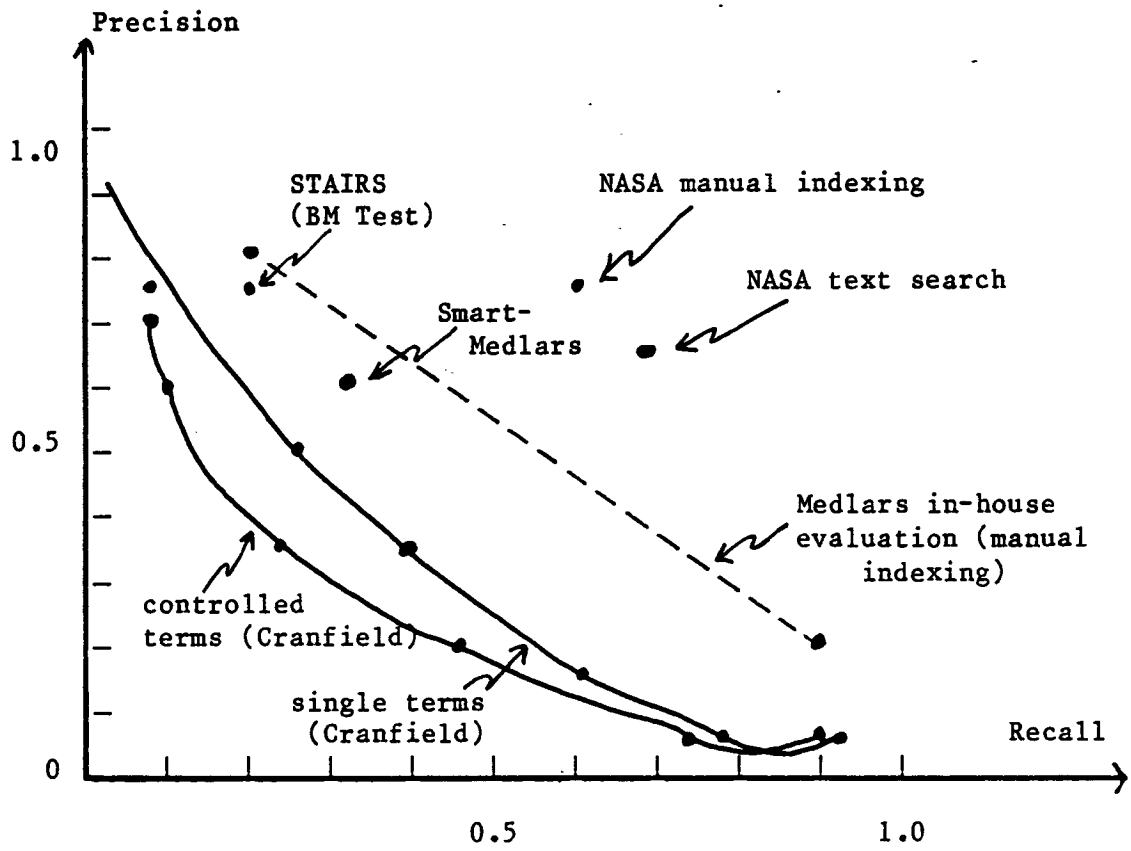
Comparative Evaluation of NASA Search System [9]
(44,000 documents, 40 queries)

Table 4



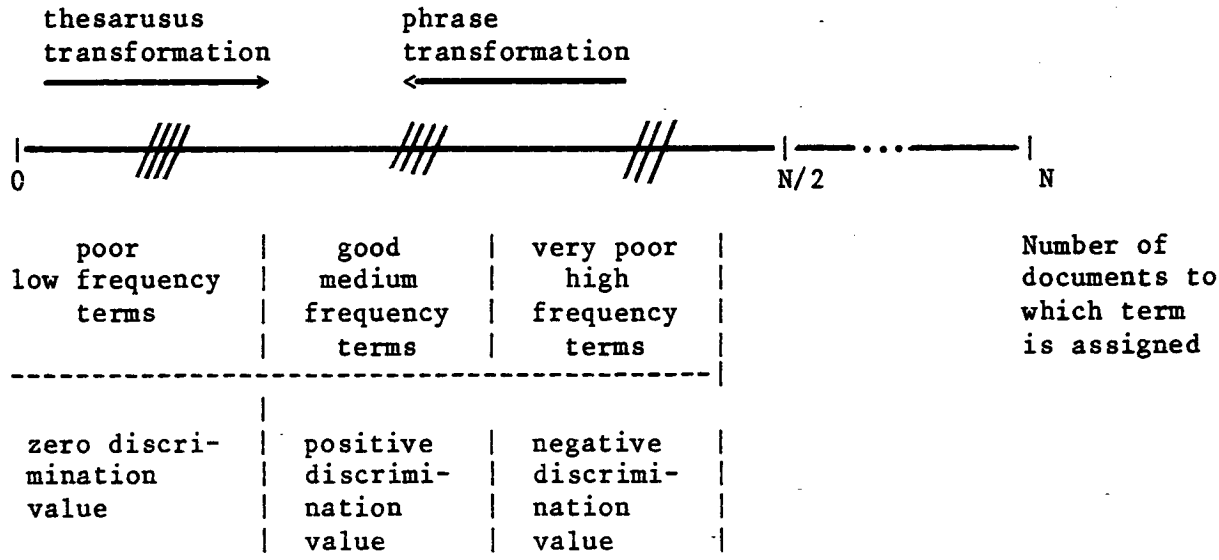
Medlars Search Service Evaluation
(adapted from [7])

Fig. 1



Comparison of Manual with Automatic Indexing

Fig. 2



Term Discrimination Model

Fig. 3