Edgar H. Sibley
Panel Editor

Evidence from available studies comparing manual and automatic text-retrieval systems does not support the conclusion that intellectual content analysis produces better results than comparable automatic systems.

ANOTHER LOOK AT AUTOMATIC TEXT-RETRIEVAL SYSTEMS

GERARD SALTON

An automatic text-retrieval system is designed to search a file of natural-language documents and retrieve certain stored items in response to queries submitted by a user. Typically, each stored item is 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 as search terms words from the text that are 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. A user interested in information about the design of small computers might formulate the query [(`minicomputers` *or* `microcomputers` *or* `hand-held calculators`) *and* (`design` *or* `construction` *or* `architecture`)]. The retrieval system would then extract, from the file, items containing the identifiers "`design`" and "`minicomputers`," or "`construction`" and "`microcomputers`." [8, 16]

The effectiveness of a retrieval system is usually evaluated in terms of a pair of measures, known as *recall* and *precision*. Recall is the proportion of relevant material actually retrieved from the file, while precision is the proportion of the retrieved material that is found to be relevant to the user's needs. In principle, a search should achieve high recall by retrieving almost everything that is relevant, while at the same time maintaining high precision by rejecting a large proportion of extraneous items. When this happens, both recall and 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 nonrelevant items tend to be obtained, but also relatively few relevant ones. That is, a very specific query formulation produces 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, thereby depressing the precision. In the latter case, one obtains high recall, but also low precision. A compromise often reached in practice is using a query formulation that is neither too narrow nor too broad. However, when a choice must be made between recall and precision, most users choose precision-oriented searches where only relatively few items are retrieved, and the user is spared the effort of examining a large amount of possibly irrelevant material—the penalty attached to a high-recall search.

In automatic retrieval systems, both query formulations and document representations can be altered to reach the desired recall and precision levels through the use of recall-enhancing devices (e.g., term truncation) to broaden the document and query identifiers, and precision-enhancing devices

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

(e.g., term weighting) to make item identifications more specific. A list of typical recall- and precision-enhancing devices appears in Table I.

Term truncation consists of using truncated terms, or word stems, instead of the original complete terms, for query or document identification. A form like "analy" would encompass the notions "analyst," "analysis," "analyzer," etc.—having a broader scope than any of the complete words. Other recall-enhancing devices involve using terms that are synonymous or related to the original ones or broader and more general. Such terms are generally available in thesauri and term hierarchies or are suggested by users during the search operations.

Term weights enhance the search precision by distinguishing the better, or more important, terms from the less important ones. Such a discrimination may also help rank the output in decreasing order of presumed importance. Other precision-oriented devices involve using term phrases instead of single terms—for example, "computer programmer" instead of "computer"—and supplying narrower or more specific terms. Useful term phrases might be available in a dictionary, or could be formed from sets of single terms that cooccur regularly in a collection of documents.

Most automatic text-retrieval systems provide for the use of truncated terms and the addition of broader, narrower, and related terms. Automatically generated term weights may also be used to distinguish items containing the more highly weighted terms from those containing terms of lower weight.

A recent article by Blair and Maron examines the well-known automatic text-retrieval system STAIRS as applied to a collection of 40,000 full-text documents—equivalent to some 350,000 pages of text—to answer 40 different user queries [1]. In STAIRS, words are normally extracted from document texts for content identification. After text words have been broadened using truncation, each word may be supplemented by lists of synonyms supplied by the user. 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 that retrieves documents in decreasing order based on total document weights, which are calculated by adding the weights of the query terms contained in each retrieved document [6].

TABLE I. Typical Recall- and Precision-Enhancing Devices

Recall-enhancing devices (term broadening)	Precision-enhancing devices (term narrowing)
Term truncation (suffix removal)	Term weighting
Addition of synonyms	Addition of 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)

Although some features of the STAIRS system are not as attractive as they might be (e.g., 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 STAIRS retrieval test conducted by Blair and Maron, an average precision value of about 75 percent (0.75) was obtained, and an average recall value of 20 percent (0.20). That is, for each of the 40 test searches, three out of four 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 were retrieved.

In this article, we will argue that not only is this level of performance typical of what is achievable in existing, operational retrieval environments, but that it actually represents a *high order* of retrieval effectiveness. We will present some major experiments comparing automatic retrieval with manual, controlled vocabulary systems on large document collections. We then address the theories underlying automatic indexing and propose a basic blueprint for implementing effective automatic retrieval systems, emphasizing that the future lies in automatic and not in manual systems.

THE BLAIR AND MARON RETRIEVAL TEST

In the Blair and Maron test of the STAIRS system, searchers were able to extract from a large collection of 40,000 documents a substantial number of useful items; since only one of four retrieved items proved extraneous, the time consumed considering useless items must have been comparatively small. However, the searchers in the Blair and Maron test were lawyers and the materials being searched were legal documents, and because the Anglo-American legal system is based on the concepts of common law and judicial precedence, many lawyers are of necessity high-recall users. In this tradition, knowing how a particular legal case must be approached often means examining all possible previous cases that may be similar in some respect to the current case. The high-precision output obtained by Blair and Maron, which rejected most nonrelevant materials, but also obtained only about 20 percent of the potentially useful items, might be entirely suitable in another environment (e.g., for research workers, university professors, and students). However, in the

case of the legal personnel that actually conducted the searches in the Blair and Maron test, a better recall performance was considered essential even at the cost of decreased search precision.

From their retrieval test, Blair and Maron derive three main conclusions [1]: First, they assert that, when high recall is essential in searching large collections, users cannot simply broaden the search request (as would be done experimentally for small collections) because of the problem of output overload. More specifically, they claim that, when broader search formulations are used, search precision may suffer intolerably, and users might be swamped with masses of irrelevant material. For this reason, the authors conclude that earlier test results showing the superiority of text-based retrieval over manual systems are not necessarily relevant to large, real-world collections.

Second, Blair and Maron argue that, when high recall is desired, manual indexing is preferable to full-text searching.

> . . . 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. [1]

Finally, Blair and Maron allege that full-text systems, and STAIRS in particular, are not particularly user friendly in the sense that, in their test, even trained searchers were unable to achieve adequate performance, and untrained users would presumably do even worse.

Despite the impressive precision performance of the STAIRS system in the Blair and Maron test environment, the authors conclude 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" ([1, p. 298]). This is surprising, moreover, because, in their study, no comparison was made between full-text-retrieval systems and manually indexed systems, nor between the retrieval performance of large versus small document collections. In this sense, conclusions drawn are unsupported by any data submitted to the reader—outside of the alleged poor recall performance exhibited by the STAIRS system in the legal case.

In fact, evidence abounds indicating that these conclusions may be more sentiment than fact. Specifically, the evidence from several retrieval evaluations conducted with very large document collections does not support the notion of output overload, although high recall naturally implies more retrieved items and hence more work in analyzing the output than low-recall searches. Moreover, comparisons between manual and automatic indexing systems on large document collections indicate that the automatic-text-based systems are at least competitive with, or even superior to, the systems based on intellectual indexing. Finally, there are automatic indexing systems that provide index terms that are not simply words extracted from document texts. Indeed, the automatic indexing results of Salton and Swanson [11, 20] that are cited in the Blair and Maron study were not based on the use of full document texts, but only on the analysis of document *abstracts*; the favorable results obtained in these studies on the effectiveness of automatic systems were achieved with abstracts (not full text), and therefore excessive input and verification demands were not placed on the system in these cases.

EXPERIMENTS WITH LARGE RETRIEVAL SYSTEMS

The Medlars Evaluation

In the late 1960s, Lancaster conducted an in-house study [7] of the Medlars demand search service, which is operated by the National Library of Medicine in Bethesda, Maryland, for searching biomedical literature. Medlars is based on manual, professional indexing by subject experts using a controlled indexing language described in the Mesh (Medical Subject Headings) thesaurus. After a manual indexing operation and a manual query formulation, the file search and retrieval operations are performed automatically.

The in-house evaluation of Medlars discussed in [7] involved searching a database of over 700,000 documents in biomedicine using a set of about 300 test queries. The search results varied widely; some queries performed perfectly (recall = 1.00, and precision = 1.00), whereas others retrieved no relevant

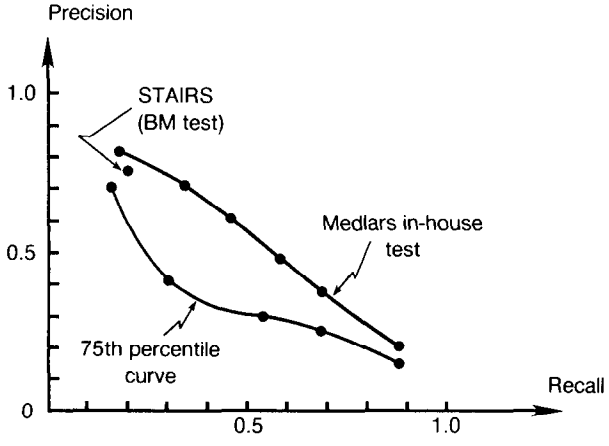

FIGURE 1. Medlars Search Service Evaluation (adapted from [7])

items at all (recall = 0, and precision = 0). For the 300 queries, the average recall performance was 0.58, and the average precision 0.50. In presenting the results, Lancaster notes 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 can be made to slide along a monotonically decreasing curve starting at the high-precision/low-recall end of the performance spectrum, and proceeding to the high-recall/low-precision end as query formulations are broadened. The resulting curve representing the performance of the Medlars search system is shown in Figure 1: A second, lower curve (also included in Figure 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 II. For the high-precision searches, the Medlars 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 level of 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. Thus, the feared output overload predicted by Blair and Maron does not occur for the Medlars search service. This is most likely not due to the manual indexing but rather to the heterogeneity of the collections, which encompass all of biomedicine and would tend to facilitate the exclusion of useless material for any one search.

The set of 500 items retrieved on average for the Medlars high-recall searches represents only seven one hundredth of a percent (0.0007) of the collection; nonetheless, such a high recall entails substantial work for the users, and only specially motivated users (e.g., lawyers) might opt to submit such broad query formulations. In [7], Lancaster remarks that

> 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?

TABLE II. Medlars Performance Points

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	0.89	0.20	500–600	135

By superimposing the performance point obtained in the Blair and Maron study of the STAIRS system—0.75 precision, 0.20 recall—on the Medlars performance curve in Figure 1, it can be seen that the STAIRS performance falls well within the range of the high-precision Medlars searches, even though no controlled language or manual indexing is used. The query broadening, recall-enhancing devices listed in Table I are available in an automatic environment like STAIRS just as they are in the Medlars controlled language environment.

The recall and precision failure analysis undertaken by Lancaster for the Medlars searches shows that manual indexing environments can also be problematic. A summary of the failure analysis for 797 recall failures (failures to retrieve relevant items) and 3038 precision failures (failures to reject nonrelevant items) appearing in Table III shows that a substantial proportion of the search failures are due to the manual indexing and the controlled language used in the Medlars environment. Some of these failures might be avoidable in an automatic indexing situation, whereas others would not. Poor search formulations and inadequate user–system interaction may occur with any retrieval system, manual or automatic. However, the conventional manual retrieval system is vulnerable in some very specific ways.

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

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

Note: Some of the failures have multiple causes accounting for totals that may exceed 100 percent.

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 thesauruses;

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. [3]

The solution Cleverdon offers is as follows:

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. [3]

Comparison of Manual and Automatic Indexing

In the mid 1970s, a comparison between automatic and manual indexing was conducted using a NASA database consisting of documents from Scientific and Technical Aerospace Reports (STAR) and 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 natural-language text-search system consisting of a machine search of document titles and abstracts, not the full text;
- a natural-language text-search system supplemented by a thesaurus of "associated concepts" prepared from the source documents;
- a controlled language indexing of the documents performed by human subject experts;
- the controlled indexing supplemented by natural-language terms extracted from the documents.

The search results for the NASA test as summarized in Table IV show that the natural-language abstract produces the best average recall for the 40 test queries (0.78) and also a high order of precision

TABLE IV. Comparative Evaluation of NASA Search System (adapted from [2]) (44,000 documents, 40 queries)

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

(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). Based on these results, it is certainly not possible to conclude that searches of natural-language abstracts are inferior, in general, to controlled language indexing. Indeed, were the NASA search population legal personnel with a recall orientation similar to the searchers involved in the Blair and Maron test, they would certainly have preferred the output produced by the automatic search system with its recall advantages of over 20 percent compared to the manual system. Cleverdon, who was in charge of the NASA test, concludes 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). [2]

The performance points for the NASA search system evaluation are plotted in Figure 2, along with the curve representing the controlled term performance for the Medlars test, and an indication for the STAIRS system. Comparing NASA and STAIRS performance on collections of comparable size shows

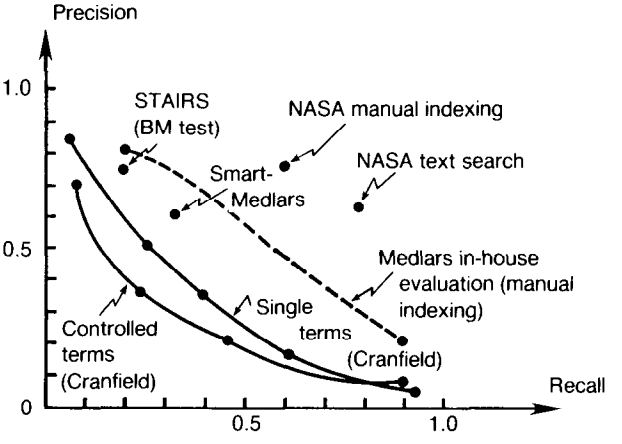

FIGURE 2. Comparison of Manual with Automatic Indexing

that the NASA searches are substantially more effective. Collection size does not seem to play an important role in search performance. Query type and homogeneity of subject matter are likely to be more important.

Many additional comparisons between automatic and controlled-term indexing systems appear in the literature. In [12], a small sample collection of 450 documents and 29 search requests is used to compare the performance of the Medlars system with an automatic indexing system based on abstract searching supplemented by the use of a thesaurus of related terms. The two systems produced almost identical results for the test collection: 0.31 recall and 0.61 precision for controlled-term indexing, versus 0.32 recall and 0.61 precision for 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 controlled-term manual indexing systems [4] as applied against a sample collection of 1400 aeronautics documents tested by 221 queries. As shown by the two typical performance curves for the Cranfield study that are included in Figure 2 [4, pp. 127 and 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 indicate that the single-term natural-language indexing provided somewhat better search results than the comparable controlled-term indexing: This is true also for the two Cranfield searches illustrated on Figure 2.

However, as mentioned earlier, 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 are available for constructing fairly sophisticated automatic document representations.

AUTOMATIC INDEXING THEORY AND PRACTICE

The effectiveness of any indexing system designed to produce useful content representations for written texts depends on two main characteristics: the *exhaustivity* of the indexing (i.e., 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 (i.e., the level of detail of a given content or index term). A high degree of exhaustivity tends to improve the recall performance of a search by permitting the identification of relevant materials that would remain unrecognized were the indexing exhaustivity lower, whereas a high degree of specificity is likely to favor search precision.

In principle, the choice of an indexing system that will be useful for content representation of natural-language texts should be based on linguistic considerations, especially semantic components. However, since linguistic analysis methods are difficult to apply efficiently to large text samples, 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 the occurrence statistics of the terms in the collection of documents. 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 is more or less inversely proportional to the number of documents to which a term is assigned [19]. Thus, terms that are assigned rarely may be assumed to be more specific than those more frequently assigned.

In judging the value of a term for purposes of content representation, two different statistical criteria come into consideration. A term appearing often in the text may be assumed to carry more importance for content representation than a more rarely occurring term, so that 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 occurs as well 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. This suggests that the specificity of a given term as applied to a given document can be measured by a combination of its frequency of occurrence inside that document (the *term frequency* or *tf*) and an inverse function of the number of documents in the collection to which it is assigned (the *inverse document frequency* or *idf*). The *idf* factor can be computed as 1 divided by the document frequency. A possible term weighting function for term i in document j [18] would then be

$$w_{ij} = tf_{ij} \times idf_i.$$

Using this term-importance definition, the best terms assigned to documents will be those occurring frequently inside particular documents but rarely on the outside. Such terms will in fact distinguish the documents of a collection from each other. Both factors of this equation are easy to calculate: The inverse document frequency of a term can be obtained in advance from a collection analysis, and term fre-

quencies can be computed from the individual documents, as needed.

The Probabilistic Retrieval Model

In the *probabilistic retrieval* model, one assumes that the most valuable documents for retrieval purposes are those whose probability of relevance to a query is largest [10, 21]. The relevance properties of the documents can be estimated by using the relevance properties of the individual terms included in the documents. Under suitably simplified assumptions, a *term relevance* weight tr_i can then be generated for term i as

$$tr_i = \log \frac{N - n_i}{n_i} + \text{constants}$$

where N is the collection size and n_i represents the number of documents in the collection with term i [5]. This formula represents the importance of the *idf* factor, since the higher the document frequency n_i of a term, the lower the relevance weight tr_i. The probabilistic retrieval model thus provides some justification for the use of the *idf* factor in the term weighting formula given on page 653, since under appropriate mathematical assumptions the idf_i factor is approximately equal to the optimal probabilistic term weight tr_i.

The Term-Discrimination Model

A different but related way of approaching the document indexing task is basing the indexing on the *term-discrimination* model [18]. Under this model, it is assumed 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 that results when a given term is assigned to the collection. The density of the document space reflects the degree to which the document representations resemble each other. This density can be measured by computing the sum of the pairwise document similarities for all pairs of documents in the collection. This means that the density of the documents will be high when the documents resemble each other a great deal (i.e., when they are indexed by many of the same terms).

Using the term-discrimination approach, the broad, high-frequency terms become the least desirable content identifiers because they will be assigned to many documents in the collection, thereby enhancing the mutual similarity of the corresponding documents. The assignment of a broad high-frequency term, because it increases the average similarity between documents, also increases the document space density. 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 preferred by the *idf* factor are also not very desirable for content identification because they are assigned to so few documents that they hardly change the space density when introduced. The very rare terms thus receive a discrimination value close to zero.

The best content identifiers will be those occurring neither too rarely nor too frequently; they will be assigned to as many as one-tenth of the items in the collection and will serve to distinguish the items to which they are assigned from the remainder. A graphic representation of the variations in term-discrimination value as a function of the document frequency of terms is given in Figure 3. 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; then, as the document frequencies become still larger,

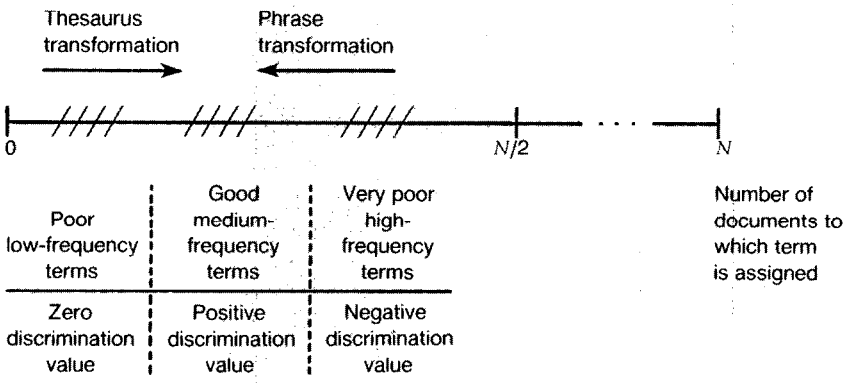

FIGURE 3. Term-Discrimination Model

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, and that terms not exhibiting the appropriate specificity should be broadened when too specific or narrowed when too broad [19]. The recall- and precision-enhancing devices included in Table I can be used for this purpose. A principal method of term broadening involves using a thesaurus, or other vocabulary grouping device, to supply synonyms and related terms of various kinds to handle the text-independent relations between terms. Term narrowing is achieved by introducing term *phrases* to replace certain broad single terms, based on a text-dependent assessment. Under the term-discrimination model the thesaurus thus assumes a specific role as a grouping device for related narrow terms. Used in this way the thesaurus and phrase transformation methods produce shifts in terms toward the center of the frequency spectrum where the content identifiers with the best specificity are located.

A BLUEPRINT FOR AUTOMATIC INDEXING

These automatic indexing strategies make possible the design of effective automatic-text-based retrieval systems that are fully competitive with conventional manual operations and can be operated without the need for human subject or domain experts for document indexing and search formulation. Summarized below is a proposed basic process [13] for automatic indexing:

- Identify the individual words occurring either in the documents or in document excerpts (e.g., titles and abstracts).
- Use a *stop list* of common function words (and, of, or, but, the, etc.) to delete from the texts the high-frequency funtion words that are insufficiently specific for content representation.
- Use a *suffix stripping* routine to reduce the remaining words to word stem form; this recall-enhancing transformation broadens the scope of the terms and can be performed automatically using a limited number of basic rules [9].
- For each remaining word stem i occurring in document j, compute a term weighting factor, which is the product of the term frequency of term i in document j multiplied by the inverse document frequency of term i in the collection as a whole. Available evaluation results indicate that term weighting improves retrieval effectiveness by distinguishing the important content terms from the less important ones [15].
- Represent each document by the chosen set of weighted word stems.

Retrieval evaluation results for this type of simple indexing for both large and small document collections indicate that even this *single-term* indexing method is competitive with, and often superior to, conventional intellectual indexing systems [2, 4, 12]. The STAIRS system used in the Blair and Maron test adheres to all these processes with the exception of term weighting. In STAIRS, term weights are assigned *after* retrieval of the documents based on term-occurrence characteristics in the retrieved document subset only; the weighting is then used to generate a *ranked* list of retrieved documents. The use of ranked document output improves the user–system interaction by alerting the user to the more important documents first; moreover, information culled from the documents retrieved early in the search can then be used to generate improved query formulations in subsequent searches.

Ideally, however, term weights should be generated before the query and document representations are compared during the search, and should be computed on the basis of the entire collection and not just a particular subset of retrieved items, which may or may not be representative of the entire collection. Certainly, terms exhibiting high-occurrence frequencies in the retrieved subset cannot be labeled effective or ineffective unless something is known a priori about their occurrence frequencies in the collection as a whole.

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

- Generate weighted word stems that are attached to the documents.
- Use a thesaurus to replace terms with low document frequencies (and near zero discrimination values) by their corresponding thesaurus class identifications.
- Use a phrase-formation process to generate term phrases that incorporte terms with high document frequencies (and negative discrimination values) based on term cooccurrences in the document excerpts.
- 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 synonymous, or related to, particular index terms. These related terms are then used automatically to expand the set of original terms. A previously available thesaurus that 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. Moreover, improved query formulations can be generated by incorporating information obtained from the texts of previously retrieved documents [13].

When search requests are submitted in Boolean form, as they are in many operational retrieval environments, weighted terms can also be incorporated: Then, an approximate, fuzzy match between the weighted term sets representing the documents and the weighted Boolean query statements can be used to produce a query-document similarity measurement that is used in turn to obtain a ranked output in decreasing order of the query-document similarity. Term weighting and output ranking are therefore available for Boolean as well as non-Boolean queries [14, 17].

CONCLUSION

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 with reference to both large and small collections indicates that properly designed text-based systems are preferable to manually indexed systems. Furthermore, as Swanson 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" [20].

REFERENCES

1. Blair, D.C., and Maron, M.E. An evaluation of retrieval effectiveness for a full-text document-retrieval system. *Commun. ACM 28*, 3 (Mar. 1985), 289–299. A recent evaluation of the IBM/STAIRS text-search system, which concludes that STAIRS does not always produce adequate search output.
2. Cleverdon, C.W. A computer evaluation of searching by controlled language and natural language in an experimental NASA data base. Rep. ESA 1/432, European Space Agency, Frascati, Italy, July 1977. A description of a large-scale test of the NASA search system using various manual and automatic text-analysis methods.
3. Cleverdon, C.W. Optimizing convenient on-line access to bibliographic databases. *Inf. Serv. Use 4* (1984), 37–47. A summary of the strengths and weaknesses of existing bibliographic retrieval systems and proposals for improving the existing methodologies.
4. Cleverdon, C.W., and Keen, E.M. *Aslib–Cranfield Research Project.* Vol. 2, *Test Results.* Cranfield Institute of Technology, Cranfield, England, 1966. The report on the most thorough evaluation of automatic versus manual text-analysis methods ever carried out, using a collection of 1400 aeronautics documents.
5. Croft, W.B., and Harper, D.J. Using probabilistic models of document retrieval without relevance information. *J. Doc. 35*, 4 (Dec. 1979), 285–295. Describes a method for using probabilistic considerations of term relevance for an initial collection search before any relevance information is available.
6. IBM World Trade Corporation. *Storage and Information Retrieval System (STAIRS)—General Information Manual.* 2nd ed. IBM Germany, Stuttgart, Germany, Apr. 1972. Contains an early description of the IBM/STAIRS system.
7. Lancaster, F.W. *Evaluation of the Medlars Demand Search Service.* National Library of Medicine, Bethesda, Md., Jan. 1968. An impressive description of an in-house test of the Medlars search system carried out at the National Library of Medicine.
8. Lancaster, F.W. *Information Retrieval Systems: Characteristics, Testing, and Evaluation.* 2nd ed. Wiley, New York, 1979. A well-known textbook in information retrieval with an emphasis on system testing and evaluation.
9. Lovins, J.B. Development of a stemming algorithm. *Mech. Transl. Comput. Linguist. 11*, 1–2 (Mar. and June 1968), 11–31. A detailed description of an automatic word-stemming algorithm.
10. Robertson, S.E., and Sparck Jones, K. Relevance weighting of search terms. *J. ASIS 27*, 3 (May–June 1976), 129–146. Describes one of the main probabilistic information-retrieval models.
11. Salton, G. Automatic text analysis. *Science 168*, 3929 (Apr. 1970), 335–343. A survey of automatic text retrieval as of 1970.
12. Salton, G. Recent studies in automatic text analysis and document retrieval. *J. ACM 20*, 2 (Apr. 1973), 258–278. An evaluation of various automatic text-analysis and indexing methods.
13. Salton, G. A blueprint for automatic indexing. *ACM SIGIR Forum 16*, 2 (Fall 1981), 22–38. A relatively nontechnical summary of an approach to automatic indexing and text analysis.
14. Salton, G. A blueprint for automatic Boolean query processing. *ACM SIGIR Forum 17*, 2 (Fall 1982), 6–25. A summary of a retrieval system based on soft Boolean logic and automatically assigned term weights.
15. Salton, G., and Lesk, M.E. Computer evaluation of indexing and text processing. *J. ACM 15*, 1 (Jan. 1968), 8–36. An early set of test results for some automatic indexing methods.
16. Salton, G., and McGill, M.J. *Introduction to Modern Information Retrieval.* McGraw-Hill, New York, 1983. A recent textbook dealing with automatic text processing and text search and retrieval.
17. Salton, G., Fox, E.A., and Wu, H. Extended Boolean information retrieval. *Commun. ACM 26*, 11 (Nov. 1983), 1022–1036. A description of a retrieval model using soft (fuzzy) Boolean logic with weighted document terms and weighted Boolean queries.
18. Salton, G., Yang, C.S., and Yu, C.T. A theory of term importance in automatic text analysis. *J. ASIS 26*, 1 (Jan.–Feb. 1975), 33–44. Contains a description of term-discrimination theory and some retrieval results based on discrimination value weighting.
19. Sparck Jones, K. A statistical interpretation of term specificity and its application in retrieval. *J. Doc. 28*, 1 (Mar. 1972), 11–21. Relates the usefulness of index terms to certain statistical term occurrence parameters.
20. Swanson, D.R. Searching natural language text by computer. *Science 132*, 3434 (Oct. 1960), 1099–1104. A pioneering small-scale test comparing an automatic text-search system with a conventional retrieval system based on manual indexing: probably the earliest result showing the superiority of automatic text searching.
21. van Rijsbergen, C.J. *Information Retrieval.* 2nd ed. Butterworths, London, England, 1979. A well-known research-oriented information-retrieval text containing many original research results, including work in probabilistic information retrieval.

CR Categories and Subject Descriptors: H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing—*indexing methods, linguistic processing, thesauri*; H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval—*retrieval models, search process*; H.3.5 [**Information Storage and Retrieval**]: On-Line Information Services; I.2.7 [**Artificial Intelligence**]: Natural-Language Processing

General Terms: Design, Experimentation, Theory

Additional Key Words and Phrases: automatic text processing, automatic text retrieval, text analysis, text searching

Received 12/85; accepted 2/86

Author's Present Address: Gerard Salton, Dept. of Computer Science, Cornell University, 405 Upson Hall, Ithaca, NY 14853-7501.