The State of Retrieval System Evaluation

Gerard Salton*

TR 91-1206
May 1991

Department of Computer Science
Cornell University
Ithaca, NY 14853-7501

*This study was supported in part by the National Science Foundation under grant IRI 89-15847.

The State of Retrieval System Evaluation

Gerard Salton*

Abstract

Substantial misgivings have been voiced over the years about the methodologies used to evaluate information retrieval procedures, and about the credibility of many of the available test results. In this note, an attempt is made to review the state of retrieval evaluation and to separate certain misgivings about the design of retrieval tests from conclusions that can legitimately be drawn from the evaluation results.

1 Introduction

The retrieval evaluation field is over thirty years old. Some early attempts at systems evaluation were made in the 1950's, but the field started in earnest with the inception of a program in systems evaluation sponsored by the National Science Foundation in the early 1960's.[1] Since that time, many dozens of retrieval evaluation studies have been conducted, and all sorts of apparently surprising results have been produced purporting to show that apparently simple-minded automatic indexing and search procedures can outperform well-established intellectual approaches to text analysis and retrieval.

Some of the available results have, however, been put in doubt by misgivings voiced by various experts about the retrieval evaluation methodologies used to conduct the studies. Eventually, the integrity of the whole evaluation field has come into question. The following quotation by Cooper is a case in point:[2]

"Another objection to the traditional (evaluation) methodology is based on experimental findings

*Department of Computer Science, Cornell University, Ithaca, NY 14853-7501. This study was supported in part by the National Science Foundation under grant IRI 89-15847.

of wide variance in relevance judgments and their sensitivity to many apparently arbitrary factors .... These and other criticisms of the traditional approach have struck at least some commentators (the present writer included) as devastating".

Many of the questions raised in the literature relate to the evaluation measures used to assess retrieval effectiveness, and especially to the use of recall and precision as evaluation parameters. More generally, the design of retrieval system tests has been criticized, including especially the methods used to obtain relevance assessments of documents with respect to information requests, and the construction of the so-called recall base (the set of documents relevant to a particular query). Finally, certain observers have objected to the extension of laboratory test results to operations of large collections functioning in realistic user environments.

Some of the questions regarding the validity of existing test results may be more reasonable than others. For example, it seems reasonable to question the claims of substantial superiority of some retrieval method over another based on the small differences in the performance measures (5 or 10 percentage points) that are often noticed in actual test situations. An attempt is made in this note to review the state of retrieval system evaluation, and to suggest approaches that provide evaluation results with some degree of reliability.

2 Retrieval Evaluation Measures

Objective retrieval evaluation measures have been used from the beginning to assess retrieval system performance. Many different parameters can in principle be used to measure retrieval system performance. Of these the best known, and the most widely used, are recall (R) and precision (P), reflecting the proportion of relevant items retrieved in answer to a search request, and the proportion of retrieved items that are relevant, respectively. The main assumption behind the use of measures such as recall and precision is that the average user is interested in retrieving large amounts of relevant materials (producing a high recall performance), while at the same time rejecting a large proportion of the extraneous items (producing high precision). These assumptions may not always be satisfied. Nevertheless, recall-precision measurements have

formed the basis for the evaluation of the better-known studies of both operational, as well as laboratory-type retrieval systems.

Recall-precision measurements have not proved universally acceptable. Objections have been raised of a theoretical as well as a practical nature. The most serious questions relate to the fact that recall, in particular, is apparently incompatible with the utility-theoretic approach to retrieval which forms the basis of a good deal of the existing information retrieval theory. Under the utility-theoretic paradigm, retrieval effectiveness is measured by determining the utility to the users of the documents retrieved in answer to a user query. Cooper has remarked in this connection that the utility and the relevance of a document represent quite different notions: a document may be relevant to a query while nevertheless proving useless for a variety of reasons. In particular, the system utility might be optimized in some cases by bringing a single relevant document to the user's attention at which point the recall value might of course be quite low.[2]

Cooper has argued that the recall measure, which depends in part on the relevance of items that have not been retrieved, may be inappropriate in all but the rare cases when a user is interested in a truly exhaustive search that catches everything that may relate to the query. In most other cases, the precision measure alone may suffice, or some other measure may replace the normal recall-precision values.

The use of recall and precision has also been criticized because for some purposes it is more difficult to deal with two parameters rather than one, and because other search parameters of potential interest, such as the collection size, and the number of items retrieved in a search, are not explicitly revealed by the recall-precision measures.[3] Many alternative retrieval evaluation measures have been introduced over the years.[4] Among the more interesting ones is Swets' E-measure [5,6] and Cooper's expected search length.[7] In the former case, the difference is measured between the probability distributions representing the query-document similarity values for relevant and nonrelevant documents, respectively. The expected search length, on the other hand, measures the average number of nonrelevant items that must be scanned by the user before retrieving a wanted number of relevant documents.

Although measures such as Swets' and Cooper's are based on sound theoretical principles, the values

obtained may be difficult to interpret in terms of actual retrieval performance in the case of the E-value, or they may be hard to compute. For example, the expected search length values depend on normally unavailable information regarding the desired number of relevant documents to be obtained in a search. Recall and precision are much easier to interpret in terms of the retrieval of wanted items and the rejection of extraneous ones, and while the recall is often difficult to generate, some users including legal personnel and patent searchers necessarily prefer high-recall results. The recall measurements may then be needed at least in those special circumstances.

Ultimately, the actual evaluation measures used to characterize the performance of retrieval systems play a relatively minor role in the total testing environment. The most solid evaluation results have been obtained with paired tests for two or more procedures carried out with otherwise fixed query and document collections. When the results of paired tests coincide for many different subject areas, document collections, and search environments, and large performance differences are obtained between different methods, the actual parameters used to measure system performance appear not to be of major concern.

3 The Retrieval Test Environment

The great majority of the retrieval evaluation studies have been conducted in a laboratory-type situation rather than in a realistic user-environment. Much of the time, the control needed to conduct reliable tests is simply not available under operational conditions, and permission to conduct potentially disturbing tests as part of the routine operations in realistic conditions should in any case be difficult to obtain.

This situation has produced much comment about the inadequacy of laboratory test designs, and conjectures about the reliability of many evaluation studies. The following critical observations may be noted, among others:

1. The search requests used for test purposes may not represent real user needs, and the relevance assessments of documents with respect to such test queries may not be representative of judgments that would actually be obtained in realistic search environments.

2. The assessment of the relevance, or utility, of a document with respect to a query must necessarily be tailored to individual users. Hence the retrieval results obtained with any given set of relevance data may be valid only for one particular set of users, rather than for arbitrary search situations.

3. The computation of the recall measures must be based on knowledge of the complete set of relevant documents with respect to each query, including both retrieved and unretrieved items. The sampling and other methods used to obtain this recall base are suspect, and as a result, unrealistic recall measurements, or recall values that favor particular search methodologies, may be produced.

4. The query and document collections used for test purposes are often unrealistically small; in such cases it is unrealistic to assume that the laboratory results are valid for realistic searches performed under operational conditions.

Some of these questions have been answered adequately in the literature. For example, retrieval experiments have been performed with distinct sets of relevance assessments produced by different user populations, and the evaluation results in terms of recall and precision have hardly changed.[8,9] The reason is that the recall and precision measurements depend primarily on the relevance, or nonrelevance, of the first few documents retrieved in response to a query, that is, those documents which are most similar to the query and exhibit the highest query-document similarities. For those items, the relevance assessments obtained from different judges are quite congruent, and the resulting evaluation data are therefore reasonable similar.

Some of the other questions raised require closer examination, and the explanations that have been offered have not overcome the feeling among some observers that the retrieval evaluation effort represents a blunt instrument with highly untrustworthy methods and results. This conclusion is reinforced by the fact that some of the large-scale tests have not produced the expected, comfortable results that would make it possible to maintain the status quo. Instead, many of the test results have been unexpected and somewhat counter-intuitive, suggesting that quite different text analysis and retrieval methods should be used in future automated retrieval situations than those currently contemplated.

Some evaluation results obtained by two well-known evaluation studies are examined in the remainder of

this note, and an attempt is made to place the retrieval evaluation effort in the proper context.

4 The Cranfield Test Environment

The Aslib-Cranfield Research Project — strictly speaking, the second Cranfield Project — was the first large-scale experiment based on paired tests of a variety of text indexing devices, conducted with fixed query and document collections.[10,11] It is also the first evaluation project that produced unexpected and potentially disturbing results. Three different indexing languages were evaluated, including single terms (similar to the keywords used in operational systems); simple concepts consisting of coordinations of single terms, equivalent to term phrases; and controlled terms extracted from the Engineer's Joint Council thesaurus. For each language various term broadening devices, such as synonym recognition, suffix removal, etc., were used, as well as term narrowing devices, such as term weight assignments, and term correlations. Thirty-three different indexing methodologies were evaluated in all, using a collection of about 300 queries and about 1400 documents in aeronautical engineering.

The main test results may be summarized by quoting from the Aslib-Cranfield report:[11]

> "Quite the most astonishing and seemingly inexplicable conclusion that arises from the project is that the single term indexing languages are superior to any other type .... This conclusion is so controversial and so unexpected that it is bound to throw considerable doubt on the methods which have been used .... A complete recheck has failed to reveal any discrepancies, and unless one is prepared to say that the whole test conception is so much at fault that the results are completely distorted, there is no course except to attempt to explain the results which seems to offend against every canon on which we were trained as librarians."

More specifically, when the 33 indexing methods are ranked in decreasing order of retrieval effectiveness using a normalized recall measurement, it is found that the seven best methods all use single term (keyword) languages; the controlled term indexing systems appear at intermediate ranks of 10, 15, 17, 18 and 19; finally, simple concept (phrase) languages provide the worst effectiveness since the 13 methodologies occupying

the last ranks from 21 to 33 are all based on simple concept indexing. In other words, in the Cranfield test environment the single term keyword languages offered the right degree of specificity for text content representation. The phrases (simple concepts) that would normally be expected to be more accurate than single terms were in fact over-specific and too narrow in scope for content identification. The controlled thesaurus entries were also not as useful as the uncontrolled keywords.

As soon as the Cranfield results because known in the middle 1960's, attempts were made to explain them away, and to find fault with the test design and the individual test methodologies. This is not surprising because to this day, many experts believe that sophisticated text analysis methods, going much beyond the keyword approach, are required to produce satisfactory retrieval results.

The most thoughtful and carefully argued criticisms of the Cranfield results are due to Swanson and Harter.[12,13] It is not possible in the present context to go into the full detail of the somewhat technical points made by these authors. In summary, a two-fold argument is made: first that the documents actually submitted to query authors for relevance assessments with respect to the corresponding queries might favor the simple (keyword) matching techniques at the expense of the more complex matching systems; and second, that large numbers of potentially relevant but nonretrieved items were never included in the evaluation so that the system was not penalized for not retrieving these items.

To explain the first argument, one must remember that it was impractical in the Cranfield environment to request relevance assessments from the query authors for all 1400 documents included in the document collection. To assess the relevance of so many document would have overburdened the users. A screening operation was therefore necessary, performed by graduate students in the Cranfield environment, to remove from the list of items to be considered for relevance most of the items that appeared to be clearly nonrelevant. The authors' relevance assessments could then be confined to promising documents only with a reasonable likelihood of being relevant.

Swanson and Harter make the point that in choosing documents likely to be relevant, some indexing methods might have been favored. For example, title searches would be favored over more general searches, if the students had chosen large proportions of documents whose titles contained one or more query words.

While the conjecture relating to title word searches appears plausible, the effect is of no consequence, because title word matches are not effective in retrieval, and did not perform well in the Cranfield environment.

The question of interest in the Cranfield test situation is the apparent superiority of single term indexing over term phrase indexing. To show any bias in the document selection process, one would need to show that documents containing single query terms would have been more likely to be termed relevant by the students, than documents containing query phrases. But that possibility appears to be counter-intuitive: the human mind is much more likely to identify and recognize document content by using phrases rather than single terms. If anything, the simple concept (phrase) searches that did not work well in practice should have been favored in the document selection process, rather than the single term searches. Swanson's conjectures that the main Cranfield results could be reversed by a less biased document selection process appear to be highly conjectural and counter-intuitive in this case.

The other main problem also raised by Swanson relates to the size of the recall base used at Cranfield. Since the search recall must be computed as the number of items retrieved that are relevant to a given query divided by the total number of relevant items in the collection, this latter quantity must be estimated as accurately as possible. In operational situations it is often not possible to capture all the potentially relevant items that remain unretrieved. In such circumstances the magnitude of the recall value could in principle be grossly overstated. In the Cranfield case, the graduate students in aeronautics who helped in the project were instrumental in identifying a portion of this recall base. In addition, a number of relevant documents that had been independently obtained by bibliographic coupling techniques were also added to the recall base. Specifically, documents exhibiting at least seven bibliographic references with other previously known relevant items were identified, and 129 of these were eventually judged to be relevant, and hence added to the recall base.

Swanson observes that the recall base used at Cranfield must have been considerably understated, because of the 129 relevant, bibliographically-coupled items that were obtained independently, only 10 had been found earlier by the graduate students in charge of the recall-base construction. Since the students were able to identify independently only 10 out of 129 relevant items found by bibliographic coupling (or 7.8 percent),

the 592 relevant documents actually included in the recall base by the students would constitute only a 7.8 percent sample of a total potential recall base of about 7600 documents.[12] Furthermore, if that many (about 7000) relevant items were actually missed, not only would the recall computations become worthless, but also the introduction of the missing items in the search process might affect different indexing methods differently, so that the ultimate ranking of the retrieval methodologies obtained in the Cranfield tests would also require revision.

While this argument appears plausible on first hearing, it is not likely to be tenable. A relationship between documents due to the sharing of bibliographic references (bibliographic coupling) represents a very special kind of relation. The set of relevant, bibliographically coupled items is thus in no way a random sample of the unknown set of relevant items. In such circumstances, the extrapolation from the proportion of identified relevant, bibliographically coupled items to the whole set of relevant is impermissible. In general, it is not credible that graduate students trained in the subject matter under consideration (aeronautics) would be able to find only about 8 percent of the existing set of items relevant to the available queries.

It is likely to be the case that the students would be unable to find all relevant items, and that some unknown (probably small) number of relevant items would be missed. That fact is, however, immaterial when paired comparisons are made between various methodologies, and a ranking is produced for the search methods of the kind obtained by Cranfield. In such circumstances, the absolute performance figures of the recall or precision are not of main interest. Instead performance is judged by using the relative performance improvement of method A over method B, and that difference does not grossly depend on the size of the recall base which remains constant over all the tests. Once again, to invalidate the Cranfield results, a bias would have to be demonstrated in the way the recall base was built, and in view of the previously given argument, such a bias is highly unlikely.

Ultimately, the Cranfield results were confirmed by the tests performed quite independently in the Smart system environment, and the objections to the Cranfield test design remain unproven conjectures that are unlikely to be of substance.

5 The Smart Test Results

The Smart system evaluation studies were initiated in the early 1960's, the aim being to show that sophisticated text analysis systems applied to query and documents texts would improve retrieval effectiveness over the standard keyword retrieval systems.[14] The earliest evaluation results appeared in 1965, and from the start, it was clear that the expectations about the power of sophisticated text analysis were not met. The use of weighted word stem indexing units extracted from document texts proved to be surprisingly powerful. On the other hand, some of the sophisticated tools that had been provided as part of the Smart system design, such as hierarchical term arrangements usable to expand the indexing vocabulary in various ways, were much less effective than expected.[15]

By the late 1960's paired experiments were performed with document collections of up to 1000 document abstracts, involving comparisons between title and abstract searches, unweighted and weighted index term assignments, word stem versus thesaurus term assignments, and so on. In several cases, the earlier findings of the manually-conducted Cranfield tests, were confirmed by these Smart evaluations: for example, title searches were again shown to be less useful than searches performed with full document abstracts; weighted index terms were generally found to be more powerful than the unweighted terms used in conventional retrieval situations; and the effectiveness of single-term indexing products held up remarkably well.[16]

Various comparison were made in the Smart environment between the better automatic indexing methods included in the Smart system and the normal, controlled-term indexing used for the Medlars search system at the National Library of Medicine.[17,18] These tests, conducted admittedly with small subcollections from the standard Medlars environment, showed that the Smart automatic indexing was not inferior to the controlled, manual indexing used at Medlars. The initial test of a Smart (single-term) word stem indexing system provided 12 percent better recall than the standard normal Medlars process, and 9 percent worse precision. Using a specially tailored thesaurus for synonym recognition and term broadening in the Smart environment provided 8 percent better recall than Medlars and 2 percent worse precision.[17]

During recent years, the Smart test collections grew in size from 1000 to about 50,000 documents, and

the evaluation methods were extended to include clustered file organizations; interactive query reformulation methods such as relevance feedback; extended, relaxed methods of Boolean logic; and many others.[19-21]

Although the Smart and Cranfield test results were complementary, the conclusions were not accepted by everyone. The main objection was related to the doubtful applicability of small laboratory tests to real retrieval operations conducted in normal user environments. This question is considered in more detail in the remainder of this note.

6 The State of Automatic Text Retrieval Systems

In a recent evaluation of the Stairs information retrieval system based on the use of 40,000 full-text documents (350,000 pages of hard-copy text) together with 40 search requests, the average search precision was reported as 0.79, while the average recall was 0.20.[22,23] Given the fact that Stairs operates in a relatively rudimentary way with unweighted keywords and Boolean query formulations, such a performance should be considered as highly satisfactory. Any time any 4 out of 5 retrieved documents are found to be relevant, the user population is bound to be pleased. The writers of the Stairs report point out, however, that the recall performance of 20 percent is unsatisfactory for many purposes, and they argue that the recall level found in the test is probably the maximum that is obtainable with full-text search systems such as Stairs.

Contradicting this, there is a commonly accepted belief that operational retrieval systems can be operated at various levels of recall and precision.[24]

> "(Operational retrieval systems) can be operated at various levels of precision versus recall. Searches can be extended almost indefinitely to achieve better and better recall, but at the cost of worse and worse precisions. As an extreme, the search could retrieve all the items (in the database) and force the user to check them all for relevance (producing very low precision, but recall equal to 1.00."

This statement is confirmed by studies where recall values of 50 percent or better have in fact been obtained under operational conditions. To cite further from McCarn and Lewis:[24]

"The evaluations provide strong evidence that users act as though they have a precision threshold; they conduct searches that retrieve a set of items at least half of which are relevant. This means that the mechanism available to the user for limiting a search to a smaller and smaller retrieval of relevant items must be more effective, the larger the database. Thus, the larger the database, the more clues to the contents of the items must be provided."

The writers of the Stairs report would probably agree, but they would argue that a query broadening operation is possible only in small laboratory systems where higher recall does not necessarily produce a disastrous overload of nonrelevant items:[23]

"on the database we studied, there were many search terms that, used by themselves, would retrieve over 10,000 documents. Search output overload is a frequent problem of full-text retrieval systems."

Hence, the argument is made that there is something fundamentally different between laboratory-type tests, where query narrowing or broadening is an option that produces various levels of recall and precision, and real-life operational systems where these operations are either unavailable or impractical. From this argument, it also follows that the accumulated test results that show recall and precision levels in the 50 percent range (such of those in [24,25]) are flawed in part because of the previously mentioned pitfalls relating to the relevance assessments and the size of the recall base, and more seriously because of the impossibility to upgrade laboratory results and render them valid under operational conditions.[26]

The question arises whether it is possible to reconcile the differing opinions concerning the usefulness of existing text retrieval systems and the validity of available test results. In considering this problem, it is necessary first of all to extend the question to text retrieval in general, rather than more narrowly to the Stairs system in particular. Blair and Maron feel that full-text retrieval systems are inherently limited: because the language is so rich and ambiguous, no user of a retrieval system can ever think of all the possible ways of expressing particular concepts, and hence the query formulations will always be inadequate and the recall results will prove unsatisfactory. As an illustration, Blair and Maron point out

that a user interested in "train accidents" might also wish to see documents dealing with "transportation system malfunctions". Furthermore many documents containing the terms "train" and "accident" may not deal with train accidents.[23]

It turns out, however, that those obvious problems are easily taken care of in modern information retrieval. It is now quite easy to provide a rich indexing profile for all documents and queries, and to introduce sophisticated text matching systems which use not only the particular words occurring in texts for the query-document similarity computations, but also the contexts in which these words occur.[27] It is thus not especially important whether a document contains the words "train" and "accident" when searching for items about train accidents. More to the point is the fact that a common context is present in all documents dealing with train accidents, and that this context is very different from the context of items dealing with "training personnel in the prevention of accidents on the job". These distinctions can be picked up in advanced full-text retrieval, and there is no obvious reason why the collection size should have much to do with retrieval system performance.

It remains to say something about the previously mentioned questions regarding the reliability of the published retrieval evaluation results. Here again, an inordinate concern with details makes critics such as Swanson, and Blair and Maron forget to larger picture. Problems do of course arise in retrieval system evaluation because of the requirement for trustworthy relevance assessments, and reliable recall-base generation. But these turn out to be second-order problems, and their effect in the end must not obscure the main aim which is to evaluate dispassionately the power of modern indexing and search methodologies.

In the Smart retrieval environment, no emphasis is placed on any of the absolute recall and precision values that may be computed. Furthermore, for at least 20 years, the Smart experiments have not been performed within a single test collection, or text environment. Instead, half a dozen different collections are normally used, comprising tens of thousands of documents and many hundreds of queries. Each collection covers a different subject area, and the queries and relevance assessments are obtained from different user populations. When all these collection environments are used in a particular evaluation study, and large-scale improvements (or large-scale deterioration) is observed for each of the test collections with some indexing

method B over a particular base case A, then it is hard to question the validity of the result by pointing to a questionable relevance assessment, or a biased searcher.

In the Smart environment, sophisticated term weighting factors are attached to query and document terms. An evaluation of several term weighting methods carried out for 5 different document collections produced on average improvement of 96 percent in average retrieval precision for a particular sophisticated term weight assignment (normalized term frequency times inverse document frequency weights) compared with the ordinary binary weights used in operational retrieval environment.[28] When the retrieval evaluation produces improvements as large as those noted for the advanced term weighting methods, and these improvements are duplicated for many other document collection environments, the conclusion seems inescapable that the binary term weights used under operational conditions should be replaced by something better.

Similar large-scale improvements in retrieval effectiveness are obtained in the Smart environment for the relevance feedback process, and for the extended system of Boolean logic that has been introduced as a replacement for ordinary Boolean query formulations. In the relevance feedback case, an average improvement of 87 percent in the average precision calculated at fixed levels of the recall was obtained for 5 different test collections with one iteration of the "dec-hi" feedback method.[29] The improvements for the extended logic using p-values of 2 over a standard Boolean logic amounted to 122 percent on average for 3 test collections.[30]

When such recent test results covering term weighting and relevance feedback are considered together with the early Cranfield and Smart results for single-term indexing, the outlines of an effective automatic text retrieval system become clear. The previously mentioned fears about the relevance assessment process and the recall-base construction will then appear to be aesthetic flaws that should be corrected, but do not invalidate the main thrust of the results obtained over the past 30 years. There should be no question at this point about the usefulness and effectiveness of properly designed automatic text retrieval systems, and about their competitiveness in relation to the more conventional approaches used in the past.

Acknowledgment

The writer is grateful to Professor W.S. Cooper for discussion and comments about many aspects of the retrieval evaluation problem.

References

1. H.L. Brownson, Research in Handling Scientific Information, *Science*, 132: 3444, December 30, 1960, 1922-1931.

2. W.S. Cooper, The Paradoxical Role of Unexamined Documents in the Evaluation of Retrieval Effectiveness, *Information Processing and Management*, 12, 1976, 367-375.

3. R.A. Fairthorne, Basic Parameters of Retrieval Tests, Proc. 1964 Annual Meeting of the Am. Documentation Institute, Spartan Books, Washington, 1964, 343-347.

4. G. Salton and M.J. McGill, *Introduction to Modern Information Retrieval*, McGraw Hill Book Co., New York 1983.

5. B.C. Brookes, The Measure of Information Retrieval Effectiveness Proposed by Swets, *Journal of Documentation*, 24:1, March 1968, 41-54.

6. J.A. Swets, Effectiveness of Information Retrieval Methods, *Am. Documentation*, 20:1, January 1969, 72-89.

7. W.S. Cooper, Expected Search Length: A Single Measure of Retrieval Effectiveness Based on the Weak Ordering Action of Retrieval Systems. *Am. Documentation*, 19:1, January 1968, 30-41.

8. M.E. Lesk and G. Salton, Relevance Assessments and Retrieval System Evaluation, *Information Storage and Retrieval*, 4, 1969, 343-359.

9. C.W. Cleverdon, Effect of Relevance Assessments on Comparative Performance of Index Languages, Cranfield College of Aeronautics, 1968.

10. C.W. Cleverdon and J. Mills, The Testing of Index Language Devices, *Aslib Proceedings*, 15:4, April 1963, 106-130.

11. C.W. Cleverdon, J. Mills, and E.M. Keen, Factors Determining the Performance of Indexing Systems, Vol. 1 - Design, Aslib Cranfield Research Project, Cranfield, England, 1966.

12. D.R. Swanson, Some Unexplained Aspects of the Cranfield Tests of Indexing Performance Factors, *Library Quarterly*, 41:3, July 1971, 223-228.

13. S.P. Harter, The Cranfield II Relevance Assessments: A Critical Evaluation, *Library Quarterly*, 41:3, July 1971, 229-243.

14. G. Salton, editor, *The Smart Retrieval System – Experiments in Automatic Document Processing*, Prentice Hall Inc., Englewood Cliffs, NJ, 1971.

15. G. Salton, The Evaluation of Automatic Retrieval Procedures – Selected Test Results Using the Smart System, *Am. Documentation*, 16:3, July 1965, 209-222.

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

17. G. Salton, A Comparison between Manual and Automatic Indexing, *Am. Documentation*, 20:1, January 1969, 61-71.dt

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

19. G. Salton, E.A. Fox, and H. Wu, Extended Boolean Information Retrieval, *Comm. ACM*, 26:11, November 1983, 1022-1036.

20. G. Salton and A. Wong, Generation and Search of Clustered Files, *ACM Transactions on Database Systems*, 3:4, December 1978, 321-346.

21. G. Salton, The Performance of Interactive Document Retrieval, *Information Processing Letters*, 1:2, July 1971, 35-41.

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

23. D.C. Blair and M.E. Maron, Full Text Information Retrieval: Further Analysis and Clarification, *Info. Proc. and Management*, 26:3, 437-447, 1990.

24. D.B. McCarn and C.M. Lewis, A Mathematical Model of Retrieval System Performance, *Journal of the ASIS*, 41:7, 1990, 495-500.

25. G. Salton, Another Look at Automatic Text-Retrieval Systems, *Communications of the ACM*, 29:7, July 1986, 648-656.

26. D.C. Blair, *Language and Representation in Information Retrieval*, Elsevier Science Publishers, Amsterdam, 1990.

27. G. Salton and C. Buckley, Flexible Text Matching in Information Retrieval, Tech. Report 90-1158, Computer Science Department, Cornell University, Ithaca, NY, Sept. 1990.

28. G. Salton and C. Buckley, Term Weighting Approaches in Automatic Text Retrieval, *Information Processing and Management*, 24:5, 1986, 516-523.

29. G. Salton and C. Buckley, Improving Retrieval Performance by Relevance Feedback, *Journal of the Am. Soc. for Info. Science*, 41:4, 1990, 288-297.

30. G. Salton, E.A. Fox, and H.Wu, Extended Boolean Information Retrieval, *Communications of the ACM*, 26:11, Nov. 1983, 1022-1036.