David C. Blair
University of Michigan, School of Business Administration

Full Text Retrieval: Evaluation and Implications

Blair, D.C.: **Full text retrieval: Evaluation and implications.** Int. Classif. 13 (1986) No. 1, p. 18–23, 14 refs.

Recently, a detailed evaluation of a large, operational full-text document retrieval system was reported in the literature. Values of Precision and Recall were estimated using traditional statistical sampling methods and blind evaluation procedures. The results of this evaluation demonstrated that the system tested was retrieving less than 20% of the relevant documents when the searchers believed it was retrieving over 75% of the relevant documents. This evaluation is described including some data not reported in the original article. Also discussed are the implications which this study has for how the subjects of documents should be represented, as well as the importance of rigorous retrieval evaluations for the furtherance of information retrieval research. (Author)

1. Introduction

Recently an evaluation of retrieval effectiveness was reported in which values of Recall and Precision were calculated for searches conducted on a Full Text document retrieval system (1), see also (2). While such studies of document retrieval systems are not uncommon, several aspects of this study make it unique. In the first place, the data base tested consisted of about 40,000 documents, while previous Recall/Precision studies based their evaluation on much smaller data bases, sometimes only a few hundred documents (3), (4). If the results of studies done on small data bases could be "scaled up" to help us understand how effective retrieval is on large data bases, then these small-scale studies would be quite informative. Unfortunately, there is reason to believe that tests of retrieval effectiveness done on small data bases do not tell us a lot about retrieval performances on large data bases (5), (6). In short, to understand how large document retrieval systems work, we must test large systems.

There have been, of course, Recall/Precision studies that have been done on large data bases (7), (8), but the results of these studies have typically suffered from unreliable techniques used to estimate Recall (specifically, the estimation of the number of unretrieved documents relevant to a particular query). Researchers such as Swanson (6), (9) have argued quite convincingly that the traditional methods of estimating Recall (e.g., having the searchers anticipate prior to their search which documents should be retrieved, or, having "experts" determine which relevant documents the searchers missed) are unreliable methods by which to calculate Recall. Swanson showed that classic Recall evaluations such as the Cranfield II project (10), (11) and Lancaster's test of MEDLARS (7), (8) used methods to Recall which were significantly biased. (Swanson estimated that approximately 90% of the relevant documents may have been missed in the Recall estimations of the Cranfield II project.)

These objections to earlier Recall/Precision studies are not meant to obscure the enormous value which these studies have contributed to the early literature of document retrieval. As Dr. Johnson once wrote, "Criticism is a study by which men grow important and formidable at very small expense (12)." What these criticisms *do* indicate, however, is that in spite of numerous Recall/Precision studies we still do not know with any accuracy how effective document retrieval is on a realistically large data base. (In her review of document retrieval tests between 1958 and 1978, Sparck Jones concluded similarly, "Overall, the impression must be of how comparatively little the non-negligible amount of work done has told us about the real nature of retrieval systems" (13).) These problems of data base size and of accurately estimating Recall were paramount in our minds when we designed our own test of retrieval effectiveness. This is why we used a realistically large data base for the test, and estimated Recall by using much more time-consuming and costly statistical sampling methods and blind evaluations of document relevance.

2. The Test Environment

The data base examined in this study consisted of just under 40,000 documents comprising over 100,000 pages of text actually stored on line. In general, the entire text of a document was not put on line, rather, some selected portion of each document was included in the data base in lieu of the entire text. The selected portions of the documents that were entered onto the data base comprised, in the judgment of the editors, the most significant or representative portions of the individual documents. As a result, the 40,000 documents on the data base represented approximately 350,000 pages of hard-copy text.

The data base itself was for use in the defense of a large corporate law suit, and access to the information was provided by IBM's STAIRS/TLS software **ST**orage **A**nd **I**nformation **R**etrieval **S**ystem/**T**hesaurus **L**inguistic **S**ystem). STAIRS software represents state-of-the-art commercial software in full-text retrieval. It provides facilities for retrieving text where specified words appear singly or in complex Boolean combinations. An inquirer can specify retrieval of text where words appear together anywhere in the document, within the same paragraph, within the same sentence, or adjacent to each other (as in "New" adjacent "York"). Retrieval can also be performed on fields other than the text of the document, such as: author, date, and document number. STAIRS also provides ranking functions which could be used to order retrieved sets of 200 or fewer documents. These functions permit the inquirer to order retrieved sets in ascending or descending numerical (e.g., dates) or alphabetic (e.g., authors) order. In addition, retrieved sets of fewer than 200 documents could be ordered by the frequency in which specified search

terms occurred in the retrieved documents. The Thesaurus Linguistic System provides the facilities for the system designer to manually create a thesaurus which could be invoked by an inquirer to semantically broaden his searches. The TLS provides the tools for the designer to specify such semantic relationships between search terms as "narrower than", "broader than", "related to", "synonymous with", and automatic phrase decomposition.

3. The Experimental Protocol

We wanted to test how well STAIRS could be used to retrieve *all* and *only* the documents relevant to a given query. In essence, we wanted to determine the values of Recall (percentage of relevant documents retrieved), and Precision (percentage of retrieved documents that are relevant). While Precision is an important measure of retrieval effectiveness, it is meaningless unless compared to the level of recall desired by the inquirers. In this case, the lawyers who were to use the system for litigation support stipulated that they must be able to retrieve 75% of all the documents relevant to a given query, and 100% of those documents they regarded as "vital" to the defense of the case (the lawyers, as was their custom, evaluated retrieved documents as "vital", "satisfactory", "marginally relevant", or "irrelevant".

4. Conduct of the Test

For the test we attempted to have the retrieval system used in the same manner it would have been during actual litigation. Two lawyers, who were the principal defense attorneys in the suit, participated in the experiment. They generated a total of 50 different information requests, and these requests were translated into formal queries by either of two paralegals, both of whom were familiar with the case and experienced with STAIRS. The paralegals would search on the data base until they found what they considered a set of documents which would satisfy the lawyers' original request. The original hard copies of these documents were retrieved from files, and xerox copies of them were sent to the lawyer who originated the request. The lawyer would then evaluate the retrieved documents ranking them according to whether they were "vital", "satisfactory", "marginally relevant", or "irrelevant" to their original information request. The lawyer would then make an overall judgment concerning the retrieved set he had received, stating whether he wanted further refinement of the query and further searching for relevant documents. His reasons for any subsequent query revisions were made in writing and were fully recorded. The information request and query formulation procedures were considered to be complete only when the lawyer stated in writing that he was satisfied with the search results for that particular query (i.e., in his estimation he had more than 75% of the relevant documents). It was only at this point that the experimenters could begin the task of measuring Precision and Recall. (A diagram of the information request procedure is given in Figure 1.) It is important to emphasize that the lawyers and paralegals

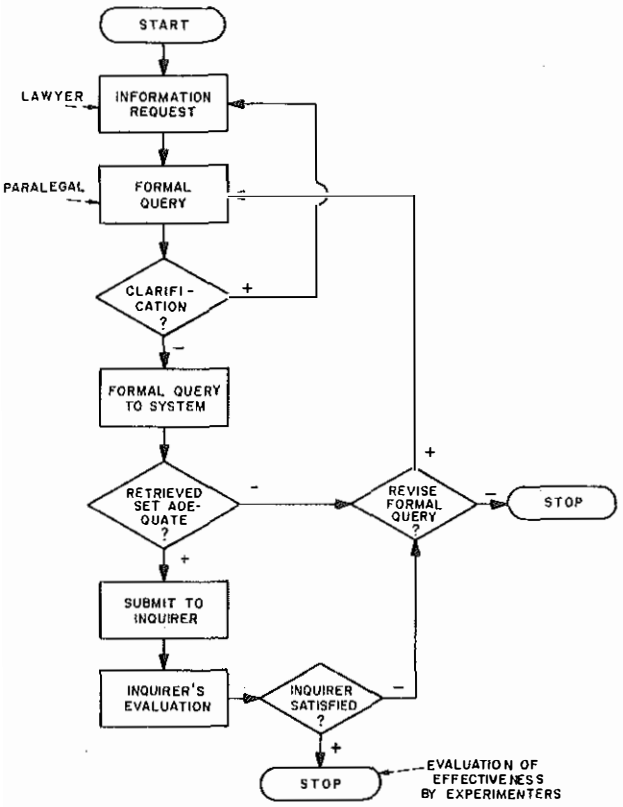

Figure 1

were permitted as much interaction as they thought necessary to insure highly effective retrieval. The paralegals could seek clarification of the lawyers' information request in as much detail and as often as they desired. The lawyers were encouraged to continue requesting information from the data base until they were satisfied that they had enough information to defend the lawsuit on that particular issue (query). In the conduct of the experiment every query required a significant number of revisions, and the lawyers were not generally satisfied until many retrieved sets were generated and evaluated.

Precision was calculated by dividing the total number of relevant documents (as judged by the lawyers) retrieved by the total number of retrieved documents. If two or more retrieved sets were generated before the lawyer was satisfied with the results of the search, then the retrieved set considered for calculating Precision was computed as the *union* of all retrieved sets generated for that information request (documents which appeared in more than one retrieved set were, of course, automatically excluded from all but one set).

Recall was considerably more difficult to calculate since it required us to find relevant documents that had not been retrieved in the course of the lawyers' searches. To find these unretrieved relevant documents we developed sample frames of subsets of the unretrieved data base which were believed to be rich in relevant documents (from which duplicates of retrieved relevant documents had been excluded). Random samples were taken from these subsets and these samples were examined by the lawyers in a blind evaluation (i.e., the lawyers were not aware that they were evaluating sample sets rather than retrieved sets they had person-

ally generated). The total number of relevant documents that existed *in these subsets* could then be estimated. Of course, no extrapolation could be made to the entire data base from these calculations, but the estimation of the number of relevant unretrieved documents in these subsets of the data base would give us a *maximum* value for Recall for each information request.

5. Test Results

Of the 51 retrieval requests processed, values of Precision and Recall were calculated for 40. The other 11 requests were used to check our sampling techniques and control for possible bias in the evaluation of retrieved and sample sets.

Table 1 shows the values of Precision and Recall for each of the 40 information requests mentioned above. In making these calculations, a relevant document was any document judged by the lawyer as being either "vital", "satisfactory", or "marginally relevant." The values of Precision ranged from a maximum of 100.0 percent to a minimum of 19.6 percent. The unweighted average value of Precision turned out to be 79.0 percent (the weighted average was 75.5). This meant that, on the average, 79 out of every 100 documents retrieved using STAIRS were judged to be relevant.

The values of Recall ranged from a maximum of 78.7 percent to a minimum of 2.8 percent. The unweighted average value of Recall was 20.0 percent (the weighted average value was 20.26). This meant that, on the average, STAIRS could be used to retrieve only 20% of those documents that would be judged relevant when the inquirers believed that they were retrieving a much higher percentage of the relevant documents (the lawyers believed they were retrieving over 75% of the relevant documents at the time).

When we plot the value of Precision against the corresponding value of Recall for each of the 40 information requests, we get the scatter diagram shown in Figure 2. Although this scatter diagram does not contain any more data than is contained in Table 1, it does reveal the relationships in a more explicit manner. We can see, for example, a heavy clustering of points in the lower right corner. This shows that in over 50% of the cases we get values of Precision above 80% with Recall at or below 20%. Looking at the lower portion of the scatter diagram we see a clustering of points showing that in 80% of the information requests the value of Recall was at or below 20%.

6. Other Findings

Several other statistical calculations were carried out after the initial Recall/Precision estimations in the hope that additional inferences could be made about the retrieval effectiveness of STAIRS. First, the results of the experiment were broken down according to each lawyer in an attempt to establish whether one of them was, prima facie, better able to use STAIRS to retrieve documents. The results were:

	Recall	*Precision*
Lawyer 1	22.7	76.0
Lawyer 2	18.0	81.4

Information Request Number	*Recall*	*Precision*
1	*	*
2	45.5%	92.6%
3	*	*
4	*	*
5	*	*
6	8.9	60.0
7	20.6	64.7
8	43.9	88.8
9	13.3	48.9
10	10.4	96.8
11	12.8	100.0
12	9.6	84.2
13	15.1	85.0
14	78.7	99.0
15	*	*
16	*	*
17	*	*
18	13.0	38.0
19	15.8	42.1
20	19.4	68.9
21	41.0	33.8
22	22.2	94.8
23	2.8	100.0
24	*	*
25	13.0	94.0
26	7.2%	95.0%
27	50.0	42.6
28	50.0	19.6
29	*	*
30	7.0	100.0
31	*	*
32	12.5	100.0
33	18.2	79.5
34	14.1	45.1
35	*	*
36	4.2	33.3
37	15.9	81.8
38	24.7	68.3
39	18.5	83.3
40	4.1	100.0
41	18.3	96.9
42	45.4	91.0
43	18.9	100.0
44	10.6	100.0
45	20.3	94.0
46	11.0	85.7
47	13.4	100.0
48	13.7	87.5
49	17.4	87.8
50	13.5	75.7
51	4.7	100.0

Average Recall = 20.0% ←——— Standard Deviation = 15.9)
Average Precision = 79.0% ←——— (Standard Deviation = 23.2)

Table 1

While there does seem to be some difference between the results for each lawyer, the variance is not statistically significant at the .05 level. Although this is a very limited test, we can conclude that at least for this experiment the results were independent of the particular inquirer involved.

Another area of interest concerned the revisions made to information requests when the lawyer was not completely satisfied with the initial retrieved sets of

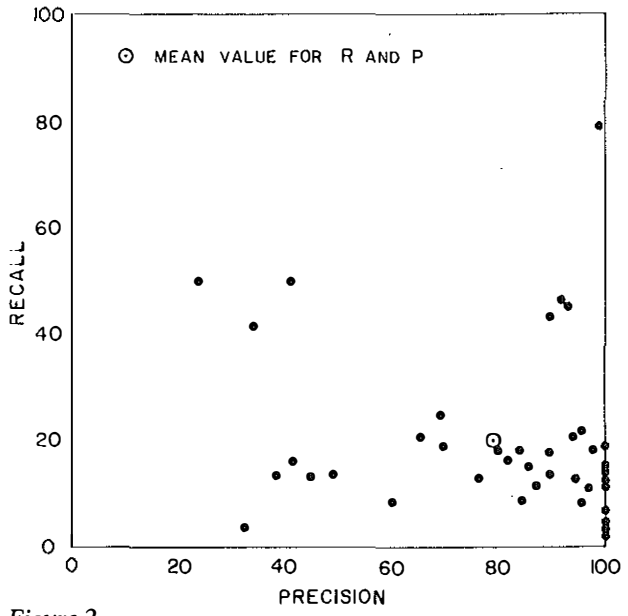

Figure 2

documents. We hypothesized that if the values of Recall and Precision for the requests, where substantial revisions had to be made (about 30% of the total) were significantly different from the overall mean values, we might be able to infer something about the requesting procedure. Unfortunately, the values for Recall and Precision (23.9% and 62.1% respectively) for the substantially revised queries did not indicate a statistically significant variance from the mean overall values for Recall and Precision.

We also tested the hypothesis that extremely high values of Precision for the retrieved sets would correlate directly with the lawyer's judgement of satisfaction with that set of documents. In other words, the lawyers, in their searches, were confusing Precision with Recall. If the lawyers *were* confusing Precision with Recall, then they would not request further searching to be done if the initial retrieved set of documents had a Precision level higher than the desired Recall levels (75 percent). But several of the information requests for which further searching was requested had initial Precision levels of over 75 percent (up to 100 percent). Other information requests began by retrieving sets of documents with high levels of Precision, but, as further searching was done, the levels of Precision dropped dramatically. Nevertheless, for each of these information requests, the inquirer eventually expressed satisfaction with the search (in spite of the degrading Precision levels). Finally, if we look at the mean Precision values for those information requests where no further searching was requested, we find a value of 85 percent. Now, while this value is higher than the mean Precision level for all requests, the difference in these two means is not statistically significant at the .05 level. Thus, it seems that the reason for the inquirers' poor Recall estimation is more complex than just a confusion of Recall and Precision.

7. Retrieval Effectiveness: Lawyers vs. Paralegals

Consider the following argument: Because STAIRS is a high speed, online, interactive system, the searcher at the terminal can quickly and effectively evaluate the output of STAIRS during the query modification process. Therefore, the retrieval effectiveness can be significantly improved if the person who originated the information request was himself doing the searching at the terminal. This means that if a lawyer worked directly on query formulation and query modification at the STAIRS terminal, rather than using the paralegal as an intermediary, the retrieval effectiveness would be improved.

We tested this conjecture by comparing the retrieval effectiveness of the lawyer versus the retrieval effectiveness of the paralegal on the *same* information request. We selected (at random) five information requests for which the searches had already been completed by the paralegal, retrieved sets had been evaluated by the lawyer, and values of Recall had been computed. (Neither the lawyer who made the relevance judgements of retrieved sets nor the paralegal knew the Recall figures for these requests.) We invited the lawyer to use STAIRS directly to access the data base, and we gave him copies of his original information requests. He "translated" these information requests into formal queries, evaluated the text displayed on the video screen, modified the queries as he saw fit, and decided when to finally terminate the search. We knew which documents he had previously judged relevant, and we had previously estimated (for each of the five information requests) the minimum number of relevant documents in the entire file. Therefore, we were able to compute for the lawyer (as we had already done for the paralegal) the values of Recall. Thus, if it were true that STAIRS would give better results when the lawyers themselves work at the terminal, then the values of Recall should be significantly higher than the values of Recall when the paralegals did the searching. The results were:

Request Number	*Recall (Paralegal)*	*Recall (Lawyer)*
1	7.2%	6.6%
2	19.4%	10.3%
3	4.2%	26.4%
4	4.1%	7.4%
5	18.9%	25.3%
Mean	10.7% (s.d. = 7.65)	15.2% (s.d. = 9.83)

Although there is a marked improvement in the lawyer's Recall for information requests 3, 4 and 5, and in the average Recall for all 5 information requests, the improvement is not statistically significant at the .05 level ($z = 0.81$). Hence, we cannot reject the hypothesis that both the lawyer and the paralegal get the same results for Recall.

8. Compnting Precision and Recall at Different Levels

When the lawyers evaluated the retrieved sets of documents, they made their relevance judgements on a scale of four. Either a document was irrelevant or else it was judged to be Marginal (M), Satisfactory (S) or Vital (V). This three-way division of relevance (M, S, and V) was

suggested by the lawyers themselves and reflects the way that they normally evaluate information of this type.

Given this three-way breakdown of relevance judgements, we can compute different values of Precision and Recall by setting different threshold levels separating the irrelevant and the relevant documents. In Table 1 the relevant documents were those judged to be either V, S, or M. In Table 2, we have calculated values for Recall and Precision for the documents judged "Vital" and "Satisfactory" (V + S) and for just those judged "Vital" (V).

Document Categories Considered Relevant	*Recall*	*Confidence (95%)*	*Precision*
1. V + S + M	20.0%	± 4.9	79.0%
2. V + S	25.3%	± 6.6	56.6%
3. V	48.2%	±14.8	18.2%

Table 2

The first thing we notice is that a Recall goes up (from 20.0% to 48.2%), Precision goes down (from 79.0% to 18.2%). This inverse relationship between Recall and Precision is well known in the field of Information Retrieval and reflects the rough trade-off that exists between the two variables; viz., for retrieval systems in general, the greater the average Precision, the lower the average Recall (and vice versa).

In looking at the values for Recall at levels 1–3, we find that they increase noticeably. This, too, should be expected in most retrieval systems. It appears that STAIRS becomes increasingly more responsive to retrieving "Vital" documents as compared to "Satisfactory" and "Marginal" ones. But is this really the case?

While there is a noticeable difference between the values of Recall at levels 1 and 2 (V + S + M and V + S, respectively) the difference is not statistically significant at the .05 level. This means that STAIRS is not significantly better at retrieving "Satisfactory" documents than it is at retrieving "Marginal" documents. At level 3 ("Vital"), however, the difference between this value of Recall and the two others is significant at the .01 level. In other words, STAIRS *does* do a better job retrieving "Vital" documents than it does retrieving "Satisfactory" or "Marginal" ones. This is encouraging. But how do these different levels of effectiveness compare with the standards expressed by the lawyers? The lawyers maintained that the minimum acceptable Recall for all relevant documents was 75%, and the 20% value we calculated for level 1 is clearly below this figure. For "Vital" documents, the lawyers stated that they must have 100% of these documents for each request. In light of this standard, the value of 48% for level 3, while it is better than level 1 or 2, is still well below the minimum standard set by the lawyers and indicates that, on the average, the lawyers would get slightly less than half of the "Vital" documents relevant to their information needs. This, perhaps, is the most crucial test of STAIRS. The "Vital" documents are those that a searcher wants most urgently. They are those that a lawyer may feel are absolutely essential to his understanding (and defense or prosecution) of some key issue. To prepare for a major trial using a system that retrieves less than 50% of the Vital documents would put a lawyer in a very risky situation.

One other important comment must be made on the method of calculating Recall for "Vital" documents alone (level 3). For 13 of the Information Requests no "Vital" documents could be found either in the retrieved sets or in any of the sample sets. If we could assume that in the case of every Information Request there must be at least one "Vital" document on the data base, then for these requests Recall would be zero. This is not an unreasonable assumption, but it is an assumption nonetheless, and we felt that it was best to give STAIRS the benefit of the doubt so we excluded these 13 queries from our calculation of Recall for "Vital" documents. If we did assume that there must be at least one "Vital" document relevant to each request, and that Recall = 0 for these 13 requests, then the average Recall for the "Vital" documents would go from 48.2% to 32.6%. This is a marked drop in estimated retrieval effectiveness. But there is an even more important consequence. In a statistical sense, this value of 32.6% is no longer significantly different from the Recall values for levels 1 and 2. If this were the case, it would mean that STAIRS does not do significantly better retrieving "Vital" documents than it does retrieving "Satisfactory" or "Marginal" ones.

9. Discussion

The realization that simple full-text retrieval can be used to identify only one out of five relevant documents as the result of a typical search may surprise those who have used such a system or had it demonstrated to them. This is because they probably will have seen only the retrieved set of documents and not the total corpus of relevant documents; that is, they have seen that the proportion of relevant documents in the retrieved set (i.e. Precision) is quite good. But they will probably not have any reliable estimate of how many relevant documents remain unretrieved in the data base. It could be argued that the results of our study are not, in fact, generalizable, and that they accurately represent the retrieval performance of only the data base we examined. I do not think that this is the case. Simple full-text retrieval is based on the assumption that it is a relatively simple matter for searchers to predict the exact words and phrases that are used in the documents which they would regard as relevant to their request, and, in fact, it is rather easy to do this. The problem is that most of the words and phrases which an inquirer would anticipate being in relevant documents would also be in many non-relevant documents. This is what causes what we referred to as "output overload" [(1) p. 296]. Identifying the exact words and phrases which are likely to appear in relevant documents is not the optimal strategy for retrieving information on full-text retrieval systems. What the inquirer must do is to successfully predict the words and phrases that not only appear in the relevant documents, but *do not* appear in the text of non-relevant documents. But this is an impossibly difficult strategy given the in-

herent flexibility and creativity of natural language [see 1, p. 295–296] for examples of how the flexibility and creativity of natural language inhibite effective retrieval).

Another important issue concerns what we can infer from this study about how documents should be represented for effective retrieval from realistically large data bases. In the first place, the study should put to rest any lingering belief in the potential for simple full-text retrieval as a method for gaining high Recall and tolerable Precision in searching large document data bases. A more subtle inference is that document representation (or subject indexing) based on vocabulary extracted from the documents to be represented may not be the best, or most complete way of representing those documents for retrieval. Since 80% of the relevant documents were missed by the searchers using the retrieval system we studied, it is clear that the vast majority of relevant documents did not contain the words and phrases used in the original search queries, in spite of the fact that these unretrieved relevant documents talked about the subjects the searchers were interested in. Since automatic indexing techniques are based almost exclusively on extracted vocabularies, it is not at all clear how effective these techniques will ever be (regardless of how complex they are) for providing high quality representations of the subject content of documents. I am not, of course, saying that automatic indexing procedures are doomed to failure. I am merely saying that since simple full-text retrieval has been demonstrated to work much more poorly than had been previously reported [see Salton (4) and Swanson (3) and the discussion of these two experiments in (1), pp. 297–298] we must anticipate that those automatic indexing procedures based on extracted vocabularies may also be less promising than originally anticipated (see also 14).

Finally, one last inference may be drawn from our study of retrieval effectiveness, and that is that we cannot draw any reliable conclusions about how effective a given information retrieval strategy or document indexing system is until we evaluate its performance using a realistically large data base, rigorously controlled experimental design, and accepted tests of statistical significance to evaluate the results. Without such a reliable experimental protocol the results of a retrieval effectiveness study can be only tentative, at best. It is an unfortunate characteristic of the field of document retrieval that almost any retrieval procedure or indexing strategy can be made to perform reasonably well on a small data base under conditions that are not carefully controlled. Sadly, the cost of rigorous studies on realistic data bases is exorbitant (our evaluation cost close to a half a million dollars), which indicates that reliable advances in our understanding of document retrieval techniques will be slow in coming.

Acknowledgement: I would like to thank Barbara Kerekes Blair for making the drawings which accompany the text.

References:

(1) Blair, D.C., Maron, M.E.: An evaluation of retrieval effectiveness for a full-text document-retrieval system. Comm.ACM 28 (1985) No. 3, p. 289–299
(2) Blair, D.C.: Searching biases in large interactive document retrieval systems. J. ASIS 31 (1980) No. 4, p. 271–277
(3) Swanson, D.R.: Searching natural language text by computer. Science 132 (1960) No. 3434, p. 1099–1104
(4) Salton, G.: Automatic text analysis. Science 168 (1970) No. 3929, p. 335–343
(5) Resnikoff, H.L.: The natural need for research in information science. STI Issues and Options Workshop. House Subcommittee on Science, Research and Technology. Washington, DC, Nov. 3, 1978
(6) Swanson, D.R.: Information retrieval as a trial and error process. Libr. Quarterly 47 (1977) No. 2, p. 128–148
(7) Lancaster, F.W.: Evaluation of the MEDLARS demand search service. Washington, DC: Natl. Libr. of Medicine 1968
(8) Lancaster, F.W.: MEDLARS: Report on the evaluation of its operating efficiency. Amer. Doc. 20 (1969) p. 119–142
(9) Swanson, D.R.: Some unexplained aspects of the Cranfield tests of indexing performance factors. Libr. Quart. 41 (1971) p. 223–228
(10) Cleverdon, C., Mills, J., Keen, M.: ASLIB Cranfield Research Project: Factors determining the performance of indexing systems. Vol. 1. Test design. Cranfield, Bedfordshire: College of Aeronautics 1966
(11) Cleverdon, C.: Cranfield tests on index languages devices. Aslib Proc. 19 (1967) p. 173–193
(12) Johnson, S.: The Idler: number 60. (Saturday, 7 June 1759). In: Samuel Johnson: Selected Writings. New York: Penguin Books 1968
(13) Sparck Jones, K.: Retrieval system tests 1958–1978. In: Information Retrieval Experiment. London: Butterworths 1981
(14) Gesellschaft für Klassifikation e.V.: Free text in information systems. Capabilities and limitations. Int. Classif. 12 (1985) No. 2, p. 95–98

Address:
Prof. D.C. Blair, The University of Michigan
School of Business Administration. Computer & Inform. Syst.
Ann Arbor, MI 48109–1234. USA