Information Processing & Management Vol. 26, No. 3, pp. 437–447, 1990
Printed in Great Britain.

0306-4573/90 $3.00 + .00

OPINION PAPER

FULL-TEXT INFORMATION RETRIEVAL: FURTHER ANALYSIS AND CLARIFICATION

DAVID C. BLAIR
Graduate School of Business Administration, The University of Michigan, Ann Arbor, MI 48109, U.S.A.

and

M.E. MARON
School of Library and Information Studies, The University of California at Berkeley, Berkeley, CA, U.S.A.

(*Received* 24 *February* 1989; *accepted in final form* 5 *May* 1989)

Abstract—In 1985, an article by Blair and Maron described a detailed evaluation of the effectiveness of an operational full text retrieval system used to support the defense of a large corporate lawsuit. The following year Salton published an article which called into question the conclusions of the 1985 study. The following article briefly reviews the initial study, replies to the objections raised by the second article, and clarifies several confusions and misunderstandings about the 1985 study.

1. INTRODUCTION

> Where there is much desire to learn, there of necessity will be much arguing, much writing, many opinions; for opinion in good men is but knowledge in the making.—Milton

In 1985 Blair and Maron published a study in *Communications of the ACM* (hereafter referred to as the "STAIRS study" which described a large-scale experiment aimed at evaluating the retrieval effectiveness of a full-text search and retrieval system. The following year Gerard Salton published an article, also in *Communications of the ACM* (1986), in which he called into question the conclusions of the STAIRS study.

There are three principal reasons why it is important to respond fully to the Salton paper, and these reasons go beyond the "mere" fact that clear penetrating discussion, critical argument and counterargument are necessary for the healthy growth of a scientific field such as ours. In the first place full-text information retrieval is a very active and important sub-field of information retrieval and it is growing very rapidly under the driving force of a dramatically improving information technology. (Simple full-text retrieval systems command by far the greatest market share of new, large-scale document retrieval systems. As of May 1988 simple full-text retrieval systems from the ten major vendors had licensed 9,375 installations (Verity Inc., 1988).) Clearly, the promise of truly effective full-text retrieval is a powerful incentive motivating work in this field.

Secondly, the STAIRS study of full-text retrieval was a milestone in the history of experimental studies of information retrieval in terms of its size, scale, setting, strict methodology and statistical rigor. The experiment ". . . took six months; involved two researchers and six support staff; and taking into account all direct and indirect expenses, cost almost half a million dollars." It is unlikely that such costly experiments on that scale will be conducted very frequently. Therefore, it is important to be very clear about its conclusions and their implications.

Thirdly, we must acknowledge the prominence of Salton as one of the pioneers in the field. For years he has been conducting experimental studies of various information retrieval techniques and his views are influential. His paper attacking the STAIRS study has already sparked much interest among researchers in information retrieval and computer science as evidenced by the early high co-citation rates for these two papers. The papers

are influencing work in connection machine based information systems (Stanfill & Kahle, 1986), intelligent information sharing systems (Malone, et. al, 1987), CD ROM (Compact Disk Read Only Memory) based retrieval systems (Zoellick, 1986), and commercial full-text retrieval systems (Berring, 1986; Dabney, 1986). In addition, an early version of the STAIRS study corroborated performance data gathered internally by WESTLAW, a large, full-text case law database, and was instrumental in causing them to begin supplementing their full-text document retrieval with search terms assigned by indexers. Therefore, it is important to respond fully to Salton's criticisms; it is essential for researchers to understand clearly what the nature and implications of the STAIRS test are.

The STAIRS study described the design, execution and analysis of a large scale, search and retrieval experiment aimed at evaluating the effectiveness of a simple full-test retrieval system. The study examined and evaluated IBMs full-text retrieval system STAIRS as used in a litigation support situation. The STAIRS database contained roughly 350,000 pages of documents which included engineering reports, internal management memos, progress reports, minutes of meetings, etc. The results of this test showed that Recall was, on average, no better than 20% with a 79% mean Precision level. Thus, no more than one in every five relevant documents in the database was retrieved—even though the lawyers using the system were convinced that, after multiple search iterations, they had in fact retrieved over 75% of the relevant documents. These conclusions about the poor Recall of the STAIRS system cannot be contested—they are the facts that the study produced. However, the study went beyond these conclusions and offered two theoretical arguments to support the view that these poor results should have surprised no one. These arguments showed why it would be difficult indeed to obtain higher Recall using a simple full-text retrieval model with a large document database.

In order for a simple full-text system to retrieve effectively, the user/searcher must be able to predict (and use as his query terms) those words, phrases and word combinations that occur in most of the relevant documents, *and which do not occur in most of the nonrelevant documents*. (See also Maron, 1988.) If a searcher can construct such a query, we shall call that an "effective query." We see that there are two interrelated parts to an effective query; predicting A, the words, word combinations, etc., that occur in the relevant documents and then B, reducing that set of terms by excluding those word or word combinations which are likely also to occur in nonrelevant documents. Let us look at these more carefully.

2. THE PROBLEM OF LANGUAGE

Consider a person who is using a document retrieval system to find information that he wants or needs for some purpose. Assume, for example, that that person is a lawyer preparing for trial, and that his litigation support system is implemented using a simple full-text retrieval system. And assume further that one of the litigation issues he must deal with concerns a certain train accident. How can he find the relevant documents among all of those that are stored on his full-text litigation support system? In thinking of the kinds of information he wants, certain terms immediately come to the mind of the lawyer in this example: It will appear obvious to him that relevant documents will, in fact, contain those "obvious" terms. Therefore, there is a good chance that he will retrieve *some* relevant documents. However, what is not so obvious to him (or to others) is that many other relevant documents (i.e., documents which discuss events leading to the accident, the accident itself or its consequences) will not contain those terms or "obvious" synonyms of those terms. How could this be? It occurs because natural language can be used to discuss a subject using an unpredictably varied and creative combination of words and phrases. (Langendoen & Postal, 1984). Just because a document is about a train accident, or discusses a train accident or deals with issues concerning a train accident, does not imply that the *words* "train" and "accident" were used in such a document. A discussion of the events surrounding a train accident can be worded in an unpredictably large number of different ways, and therefore many of these discussions may not contain the terms "train" and "accident" at all. Furthermore, an accident from one person's point of view, could be described

as an "unexpected event," or an "unfortunate occurrence" or an "untimely incident," etc. Yet, from another person's perspective the same accident could be described as a "tragic disaster" or a "monumental blunder" or a "transportation system malfunction," etc. Documents and reports about the causes and consequences of a given train accident written by different people may use a tremendous variety of different words and phrases to express similar ideas; and they can be written without using the words "train" and "accident." Such documents as those described above could be relevant (for the lawyer in our example) but not retrieved if the full-text search query contained only the terms "train" and "accident." They could be retrieved using a full-text system only if the searcher could find some not so obvious terms and phrases that might have been used to talk about the train accident. But finding such terms is a very difficult task, since it requires the searcher to imagine all of the many different ways that various authors might describe a train accident. To understand why this is so difficult to do imagine someone asking you to estimate how many times you used the words "computer" or "information" in your conversation or writing in the last week. Most, if not all, of us would not be able to give even a rough estimate of this. But this is exactly the kind of linguistic fact that simple full-text information retrieval systems require us to estimate. In fact, it is really much more difficult because these systems ask the searchers to predict the word occurrences for *other* individuals' writings—writings we may never have read, written by individuals we don't even know.

Now consider the other side of the problem of how to construct an effective query when using a full-text system—consider the set of *nonrelevant* documents. Again, continuing with our illustrative example, we must recognize that many documents might contain the terms "train" and "accident" (or their synonyms) and *not* be about the train accident question, or in fact, not about trains or accidents. They could be about how to *train* employees so that they will avoid an *accident*. Or, they could be about trains and accidents, but not the train accident we are concerned with (e.g., they could be about "accidental uncoupling of trains"). And so on. We could multiply examples endlessly. They key point here is that words such as "train" and "accident" are general purpose in the sense that they can be and are used to describe and discuss a very wide range of subjects, topics, problems, and issues *other than* train accidents. Their meaning is ambiguous and that ambiguity is resolved only in context, that is, it is resolved within the context of the other words and phrases that surround them in the texts in which they occur and within the even broader context of the activities which produced or used the documents.

It is clear from the examples reported in the STAIRS study that the number of different ways in which an author of a document could write about a particular subject was unlimited and unpredictable (see Blair & Maron, 1985, p. 295–296). The STAIRS experiment also demonstrated that the same word can be used in an unpredictable variety of different meaningful contexts in various documents (causing a relatively high incidence of "false drops"). Using the full text of a document to represent it for retrieval is like adding hundreds of marginally useful and spurious index terms which, collectively, become just so much "noise" in the system.

3. THE PROBLEM OF DATA BASE SIZE

But even a simple full-text retrieval system is tolerable for a small database of a few hundred documents, and it has been found that they are, at this size, competitive with more traditional retrieval systems. But as the database of documents grows to a more realistic size we find that the large size of a document collection exacerbates the problems of language described above. As we discussed before, the tremendous variety in natural language text causes reasonable search terms to appear not only in the meaningful contexts which the searcher wants to see, but also in an unpredictably large number of natural language contexts which do not discuss what the searcher wants. Because of this large number of spurious contexts in which words or phrases may occur the "hit" rates for terms in simple full-text retrieval systems are high. As a result, on a realistically large system, the searcher is frequently swamped with excessively large sets of retrieved documents (many of which are irrelevant)—a phenomenon we called "output overload." Since the searcher

cannot read every document in a large set of retrieved documents, he must adopt a strategy of reducing the sets of retrieved documents to a reasonable size. The most widely used strategy for reducing output overloaded is to add, conjunctively, different terms to the search query in order to retrieve only those documents which contain *all* of those query words and conditions. However, as one requests the logical intersection of the sets of those documents containing the query terms, the probability of retrieving a relevant document (or the percentage of relevant documents retrieved) drops off drastically. One can use this strategy of taking logical intersections to reduce the size of the output, but in so doing the searcher is also reducing Recall (Blair, 1980). In short, the retrieved sets of documents generated by a simple, full-text retrieval system tend to be exceptionally large. These large retrieved sets force the searcher to reduce their size by conjunctively adding more terms to his search queries—a strategy which is bound to exclude more relevant documents with the addition of each successive intersecting search term.

Another problem with retrieving documents by anticipating which words and phrases were used to discuss a topic is that often the information necessary for retrieval is not contained in the text of the document. In the STAIRS experiment we frequently found that documents often had implicit links between them in that they discussed the same issue, responded to a document making a request, made a commitment, provided information about a topic or activity, made a judgment or evaluation of another individual's proposal or statement, and so forth. For example, in the lawsuit that the STAIRS system was used to provide information control it was frequently necessary to establish evidence for "who knew *what* about the litigated issue, and *when* did he know it," or, "did anyone object to X's proposal?" To search for evidence of this type, one would have to identify every possible author of this type of document *and* describe every possible way in which such topics could be discussed, an impossible task on a document collection as large as the one studied in the STAIRS evaluation. In addition, we frequently found documents germane to the lawsuit authored by individuals who were not employees of the company engaged in the lawsuit, and who were therefore very difficult for the lawyers to identify and find the documents that they authored. There is growing evidence that some designers of information systems recognize the importance of these implicit document links and have designed systems which force the authors of documents to establish links between documents which are not explicitly described in their text. Examples are the COORDINATOR (Winograd & Flores, 1987; Flores, *et al.*, 1988), the LENS system (Malone, *et al.*, 1987) and Filenet's WORKFLO program.

One conclusion of the STAIRS study was implicit in the reported evaluation, but deserves to be made explicit. That is, the value for Recall, although low, represents a *maximum* value because it was based on estimating Recall for small subsets of the document collection, not the entire database. If we examined the entire database we probably would have found more unretrieved relevant documents. The "actual" value for Recall, if it could be calculated, would be significantly lower. It is also the case that as low as Recall was in the STAIRS study, it was probably *higher* than a typical searcher would get on another full-text retrieval system of the same type. The reason for this is that the environment in which STAIRS was tested was unusually favorable for effective retrieval. The lawyers who used the system had been working on this particular litigation for over a year and were not only intimately familiar with the issues in the complaint, but had been instrumental in supervising the selection of the documents on the database. Each of these documents was germane to at least one of the 13 issues in the complaint. The paralegals who did the searching for the lawyers had been the ones who actually selected (under the lawyers' guidance) the documents to be included in the database. In addition, they had had a great deal of training in the use of STAIRS by IBM personnel and had the continuing support of their technical staff. It's rare that a large, operational information retrieval system would be used solely by inquirers and searchers who had actually designed its logical structure and selected the documents that comprised its database. (In effect, STAIRS was being used to manage a personal document database). One would expect that inquirers and searchers such as these would be more adept at retrieval than a typical inquirer using a typical information

retrieval system. The fact that there was no "learning" curve (search results at the end of the test were not significantly better than those conducted at the beginning) in evidence for the use of the system during the test means that the searchers and inquirers were probably performing to the best of their ability. This leaves us with the ineluctable conclusion that the Recall levels which were found in the STAIRS study as relatively high, and that typical inquirers using typical systems would be likely to attain comparable levels only in the best of circumstances and at considerably greater effort.

3.1 *The Salton argument*

Professor Salton contests the results of the STAIRS study because he believes that "the future lies in automatic and not in manual systems." He may be entirely correct in this belief. The STAIRS study, which revealed very poor Recall, does not contradict him. In fact, it is clear that much can be done to significantly improve the retrieval effectiveness of commercial full-text retrieval systems. Nevertheless, in the process of arguing to support his beliefs about full-text vs. manual system, Salton has confused the results and implications of the STAIRS study.

The purpose of the STAIRS study was to conduct a very large scale, rigorously controlled, empirical test of the full-text document retrieval system in an operational setting, in order to evaluate its retrieval effectiveness when used for litigation support. It tested the performance of a basic, "bare bones," commercial full-text retrieval system—a system that was *not* enhanced by the sorts of "refinements" described by Salton. Furthermore, the STAIRS study did not examine how full-text techniques could be used to retrieve *abstracts* as opposed to complete documents. Therefore, when Professor Salton reads into the STAIRS study something more than an evaluation of a simple full-text system, as described above, then he is clearly misguided.

Before looking closely at Salton's objections to the STAIRS study, it must be pointed out that he is confused about some important fundamental aspects of the test. He mistakenly believes that "the materials being searched were legal documents." He also mistakenly believes that the lawyers, who were the users of STAIRS in the test, were dissatisfied with Recall values of 0.2 because they were doing legal precedent searching—an activity that demands high Recall. The database used in the STAIRS study consisted of technical and engineering reports, correspondence, minutes of meetings, etc., all of which were germane to a large scale corporate lawsuit. There was no legal precedent searching. No careful reading of the STAIRS study could conclude that it dealt with legal precedent searching. How Salton drew this conclusion is a mystery.

Now let us look at Salton's specific objections to the reported evaluation of STAIRS. These are as follows:

> 1. . . . the evidence from several retrieval evaluations conducted with very large document collections does not support the notion of output overload . . .
> 2. . . . 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 system based on intellectual indexing.
> 3. 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 that are cited in the Blair and Maron study were not based on the use of full document texts, but 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.

Let's consider each of these individually:

3.1.1 *Salton's first objection: No evidence for output overload on large systems.* Output overload is caused by the fact that most of the frequently used search terms will occur in a relatively large number of documents in a collection. Therefore, a search using such a term, by itself, could result in a large number of "hits" causing the searcher to be inundated with a very large number of retrieved documents. (On the database of 40,000 doc-

uments used in the STAIRS study, there were many search terms each of which occurred in over 10,000 different documents. Thus, if an inquirer were to use one of these terms by itself as a search query he would retrieve over 10,000 documents.)

The nature of the distribution of occurrences of search terms in a document database has been well known for quite some time and is based on repeated empirical verifications. The occurrences of search terms in the running text of a full-text system has been shown to follow a hyperbolic rank-frequency distribution (Zipf, 1949), and the distribution of terms in a manually indexed database is either hyperbolic (Van Rijsbergen, 1979; Little, 1963) or a closely related log-normal (Wall, 1964). These empirical studies are important to the discussion of output overload because they have demonstrated clearly that the number of occurrences (or "hits") for a given search term in a document database increases as the total number of term occurrences in the database increases. That is, as you add documents to the database the number of occurrences of individual search terms (either in full-text or manually assigned index terms) increases. This causes the problem of output overload in large databases. To deny its existence, is to deny an empirically demonstrated fact. Output overload presents difficulties in a simple full-text retrieval system because, given the nature of the Zipfian distribution, the occurrences of the most frequently appearing "content" terms (which will also be the most frequently selected search terms (Nelson, 1988)) in the database vary according to the *total number* of term occurrences in the database. (When the Zipf distribution is plotted on a log:log scale, with term frequency represented on the y-axis and the rank order of terms from most frequent to least represented on the x-axis, then the distribution will be linear with a slope of -1. This means that x and y intercepts will be equal, which in turn means that the frequency of the most frequently appearing term will be equal to the total number of unique terms (the rank number of the lowest ranking term). Therefore, as more documents are added to the database, the frequencies of the existing terms as well as the total number of unique terms will increase.) Since there are significantly more total term occurrences in a simple full-text retrieval database than in a more selectively indexed one, there will be, resultingly, much higher "hit" rates for the most frequently selected search terms in a simple full-text retrieval system. Strangely, Salton denies the existence of output overload while at the same time acknowledging this very phenomenon:

> 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.

What Salton calls ". . . a large amount of possibly irrelevant material . . ." is precisely what was meant by output overload in the STAIRS study.

The standard strategy for reducing output overload is to conjunctively combine several or many different terms in the search query in order to select just the documents which contain *all* of those query words and conditions (Blair, 1980). To make a long story short: As one requests the *logical intersection* of the sets of these documents containing the query terms, the probability of retrieving a relevant document drops off drastically. Thus, as one attempts to counteract output overload (which is a characteristic of *large* files) using this strategy, Recall drops off very quickly (Blair and Maron, 1985, pp. 296–297). On the other hand, when one is dealing with *small* collections of the size that Salton refers to in his 1970 paper (e.g., 273 articles in one set of experiments and 450 articles in the other), and *especially* when one is using a file of abstracts (as opposed to complete documents) on such small databases, there is no problem with output overload. Therefore, it is not necessary to take logical intersections of search terms in order to reduce output overload.

This was the main reason the earlier studies of Swanson (1960) and Salton (1970) on full-text retrieval were cited in the STAIRS study. Because they were using small files, they did not encounter output overload, and therefore it was not necessary to attempt to narrow the size of the output by using the standard strategy (described above) which results in low Recall. It is for these reasons, we believe, that they were unrealistically optimistic

about the future of full-text search techniques. Instead of denying that output overload is a characteristic of large files, Salton should acknowledge it and then conduct rigorously controlled, large-scale retrieval experiments to show how some of the very techniques (such as the probabalistic weighting of search terms) that he discusses in his recent *CACM* article can be used to deal with output overload.

3.1.2 *Salton's second objection: Previous Recall/Precision Studies offer evidence that ". . . automatic-text-based systems are at least competitive with, or even superior to, the systems based on intellectual indexing."* The original intention of the STAIRS study was not to offer a critique of *all* previous Recall/Precision studies conducted on computerized information retrieval systems, for the obvious reason that these earlier studies have been extensively critiqued in the literature and these critiques are simply too numerous to cite. The interested reader should direct himself to the monograph by Karen Sparck Jones (1981) which collects and summarizes the findings and discussions of these earlier Recall/Precision studies. After considering all the Recall/Precision studies and their ensuing discussions generated between the years 1958 and 1978 (covering all of the studies which Salton cites as evidence in his article) Sparck Jones remarks:

> What conclusions can be drawn about the state of information retrieval research from such a survey [as this]? More specifically, what progress has been made over the last 20 years in obtaining substantively valuable results from methodologically sound experiments?
>
> 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 . . . our ignorance is large: to take a conspicuous instance, we have virtually no information about the real recall levels of large online search systems, or about real recall for many retrieval schemes investigated by research workers.

Again, these critiques are well-known in the information retrieval literature. Also well-known is the primary reason for the problems with previous Recall/Precision studies. Sparck Jones continues:

> Conducting large test programs in document retrieval is . . . extremely laborious; it requires resources which are not available to many individual projects.
>
> It is nevertheless the case that the lack of solid results must be attributed primarily to poor methodological standards.

Traditional Recall/Precision tests of information retrieval systems have suffered from one or more of four principal methodological weaknesses:

1. using an unrealistically small database of documents;
2. not using reliable techniques for judging the relevance of unretrieved documents;
3. not conducting retrieval in a realistic, operational environment; and,
4. not using reliable tests of statistical significance to interpret the resulting data.

These difficulties have vitiated every Recall/Precision study which has been conducted prior to the STAIRS study, and have made these earlier studies not only inconclusive (as far as our understanding of Recall/Precision levels) but also incommensurable.

These four methodological problems with earlier Recall/Precision studies were major considerations during the design of the STAIRS study, and efforts to circumvent these potential problems are clearly documented in the article:

1. The STAIRS study used an operational database of realistic size (approximately 40,000 documents).
2. All relevance judgments were made by the inquirers who originated the search queries. The number of unretrieved relevant documents was estimated by using rigorously controlled statistical sampling techniques with the inquirers evaluating all the sample sets of unretrieved potentially relevant documents.

3. The lawyers (inquirers) used the retrieval system in *precisely* the same way in which they intended to use it during the defense of the lawsuit.
4. All test data were subject to standard, accepted statistical tests of significance, and many controls were maintained to test, among other things, the consistency of inquirers' relevance judgments and the effectiveness of the sample "frames."

All of this was explained in detail in the paper (see Blair & Maron, 1985, pp. 291–293), but the significance of this careful experimental methodology has, apparently, been missed. None of the Recall/Precision studies which Salton cites in his article is without some or all of the critical methodological flaws described above (see Table 1), and all of them suffer from the most crucial design flaw of these earlier experiments: demonstrably unreliable methods of estimating the number of unretrieved relevant documents. These problems with the traditional methods of estimating the number of unretrieved relevant documents have plagued Recall/Precision studies for the last thirty years and have been discussed at length in articles too numerous to cite here and passages too lengthy, in aggregate, to quote (the interested reader is directed to Sparck Jones as a first source; see especially Chapters 5, 12 and 13 *inter alia*). But to get a flavor of these critical discussions, we can look at some of what Swanson (whom Salton quotes in his article and whose opinion he clearly values) has to say about these early Recall/Precision studies which Salton bases his arguments on:

> In the Cranfield II project, an attempt was made, prior to any retrieval tests, to identify all possibly relevant documents in an experimental collection by directly examining every document with respect to each question. It has been shown [by Swanson] that this method very likely missed about 90 percent of the relevant documents . . .
>
> Lancaster's tests of MEDLARS (Medical Literature Analysis and Retrieval System) involved a method for determining "recall" (the percent of relevant documents retrieved) that could be misleading . . . The search for potentially relevant documents necessarily involved rejection of many that were judged not to be potentially relevant. But this judgment was not made by the requester, and the rejected documents were never submitted to the requester. There is no way of knowing, then, if those who were doing the screening were not systematically excluding certain classes or types of documents that might have been judged relevant by the requester . . .
>
> Using experimental collections of a few hundred documents, Salton compared MEDLARS with a fully automatic system for searching the text of abstracts and inferred that such automatic systems should replace those that depend on manual indexing. It is pointed out here that there is as yet no evidence to support an inference of this kind for very large document collections.

Table 1. Methodologies of recall precision studies

Studies cited in Salton (1986)	Characteristics of recall precision study			
	Large data base	Relevance judgments exclusively by inquirers	Realistic operational retrieval	Data subject to tests of statistical significance
MEDLARS (Lancaster, 1968)	Yes	No	Yes	No
Cranfield II (Cleverdon, 1966)	No	No	No	No
NASA data base (Cleverdon, 1977)	Yes	No	No	No
SMART-MEDLARS Comparison (Salton, 1973)	No	No	(Used same queries as MEDLARS)	Yes[1]
STAIRS (Blair & Maron, 1985)	Yes	Yes	Yes	Yes

[1]Tests of statistical significance were used to compare the retrieval data of SMART and MEDLARS, but no test was conducted to see whether the data from the SMART system alone were significant.

None of the studies cited by Salton used methods of estimating unretrieved relevant documents that were rigorous and reliable. None used relevance judgments made *exclusively* by the inquirers who submitted the original queries to the system; none used rigorous statistical sampling techniques to estimate the number of unretrieved relevant documents; and only one subjected its final data to standard tests of statistical significance. Even apart from the other methodological problems which these earlier studies fell prey to, the unreliable methods they used to estimate Recall would make their results incommensurable with the results of our evaluation of the STAIRS system. In short, none of these earlier studies can reject the hypothesis that their Recall calculations (even for small databases) are biased to an unknown extent. Again, Swanson (1977) comments:

> Salton does not make any direct claim that the results of his small-scale tests imply that SMART could perform as well as MEDLARS on the full MEDLARS data base of over 800,000 documents. He confines himself to the claim that there is no evidence to the contrary, but on this basis advocates fully automatic systems. It is no doubt true that there is no evidence to the contrary, but no tests have yet (1977) been performed, to my knowledge, which could have yielded such evidence. A simple statistical argument will show that tests on collections of a few hundred documents are not sufficiently sensitive to permit even rough estimates of performance on very large collections of hundreds of thousands of documents.

None of the previous Recall/Precision studies has produced evidence *against* automatic systems for the simple reason that they have not produced statistically significant evidence *for* the accurate estimation of Recall on any of these systems. Salton relies on the quantitative results of these early Recall/Precision studies with what Mark Twain would call, "the calm confidence of a Christian with four aces." But Salton's conclusions are, to use his own words, ". . . more sentiment than fact" (Salton, 1986). The STAIRS study of full-text retrieval has demonstrated, for the first time, in a controlled and statistically rigorous fashion, just how good these retrieval techniques are on a realistically large database used in an operational environment. Now, researchers in information retrieval know with greater clarity how effective these systems are.

3.1.3. *Salton's third objection: "There are automatic indexing systems that provide index terms that are not simply words extracted from document texts . . . the favorable results obtained in (the Salton and Swanson) 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."* It has already been shown that the earlier retrieval effectiveness studies which Salton cited in his article could not reject the hypothesis that their Recall estimations were significantly biased. To compare these earlier studies to the STAIRS study is like comparing educated guesses to the results of a rigorously controlled experiment. But apart from the incommensurable nature of these tests, there is an even more fundamental problem with this objection. Namely, the STAIRS study made no comparison, or claim, about the effectiveness of full-text retrieval versus retrieval based on abstracts of documents. To attempt to distill a negative assessment of abstract retrieval from the STAIRS study is to mix apples and oranges. Even more curious, though, is Salton's comment that because retrieval in his and Swanson's studies was based on abstracts " . . . excessive input and verification demands were not placed on the system in these cases." This is in answer, of course, to the comment in the STAIRS study that:

> . . . the full-text system incurs the additional cost of inputting and verifying *20 times* the amount of information that a manually indexed system would need to deal with. This difference alone would more than compensate for the added time needed for manual indexing and vocabulary construction.

Abstract-based systems would certainly lessen the amount of information input to a database *vis á vis* a full-text system. But there is one problem with abstract-based retrieval systems—most documents don't come with abstracts. The database we studied consisted of internal and external correspondence, memoranda, reports, engineering specifications,

minutes of meetings, etc., none of which was abstracted in any way. While it certainly would have been easier to input 40,000 abstracts onto the STAIRS database rather than 40,000 full-text documents, there were no abstracts for the documents to begin with. Is Salton recommending that the experimenters should have taken the time and effort to write an abstract for each of the 40,000 documents in order to save the time and effort of inputting the full-text of the documents? If he is, it is certainly a curious recommendation, indeed. And it is hardly an alternative that would offer an advantage over manual indexing. It also gives evidence of a rather marked misunderstanding of the commercial use of document retrieval in businesses and organizations in both the public and private sectors (the predominant market share for systems such as STAIRS). The documents which comprise databases for these organizations rarely come with abstracts in the way that articles in professional journals do. Yet these types of documents are the ones which are the primary concern for the majority of commercial document retrieval applications. (Every nuclear power plan in the United States, for example, must maintain access to, on average, over 21 million documents of this type. Only a fraction of a percent of these documents come with abstracts already written.)

4. CONCLUDING REMARKS

One of the important points of the STAIRS study which Salton has apparently missed is that the test was, in a large part, a conscious attempt to raise the standards and methodological rigor of information retrieval evaluations to a level comparable to other more established empirical disciplines. Information Retrieval research is facing a crisis right now of significant proportions. It appears that we know a fair amount about how small document retrieval systems perform (those of a thousand documents, or less), but there are very few data and almost no theory to tell us about the performance levels of large commercial systems of realistic size (see the Sparck Jones quotation, *supra*). The conclusions about the Recall levels of the few large-scale information retrieval studies which preceded the STAIRS study were vitiated by, as Sparck put it, "poor methodological standards." Yet if Information Retrieval is to be considered a science of any consequence, it must tackle the important and pressing problems of the field. The most critical problem for Information Retrieval research now is to give us an effective model for how large, operational retrieval systems work. This is precisely what the STAIRS study attempted to do.

Information Retrieval research is at a point similar to the position where Data Base Management System research was some 20 years, or so, ago. At that time, research in Data Base was in transition from the study of small-scale research-oriented systems characterized by the small size of their databases, the uniformity of their records, and their operation under ideal or unrealistically simple conditions. Imagine what the last 20 years of database research would have looked like (and how unrealistic it would have been) if it had limited its study to evaluating the performance and effectiveness of these small-scale systems. Major issues such as "concurrency control," "data integrity," "logical and physical independence," "data dictionaries," "database machines," "distributed data bases," etc., would probably not have been considered important if researchers only looked at small systems. Even problems such as "update and deletion anomalies," which inspired the early work in the normalization of relational logical structures, would only be seen as an *irritation* on a small system, not a major handicap. Interfaces between logical, storage and physical structures would never become much of an issue since, on small systems, just about any one of the interfaces available for use would work reasonably well. One could go on indefinitely giving examples of how the pressures of designing large-scale commercial DBMS's forced researchers to confront issues which researchers who worked on small systems never anticipated, or even realized were important. Information Retrieval is now at a point where it too must shift its emphasis away from the study of small, research-oriented systems and towards a more rigorous analysis of the pressing problems which large-scale commercial systems pose. If information retrieval research is successful in managing this transition, researchers can look forward to future work of a richness and complexity comparable to the recent history of database research.

REFERENCES

Arthur D. Little Inc. (1963). *Centralization and Documentation*. Cambridge, MA.

Berring, R. (1986). Full-text databases and legal research: Backing into the future. *High Technology Law Journal, 1(27)*, 27–60.

Blair, D.C. & Maron, M.E. (1985). An Evaluation of Retrieval Effectiveness for a Full-Text Document-Retrieval System. *Communications of the ACM, 28(3)*, 280–299.

Blair, D.C. & Maron, M.E. (1985). *Technical Correspondence, Communications of the ACM, 28(11)*, 1238–1242.

Blair, D.C. (1980). Searching biases in large, interactive document retrieval systems. *Journal of the American Society for Information Science, 31*, 271–277.

Cleverdon, C.W. & Keen, E.M. (1966). *Aslib-Cranfield Research Project*, Vol. 2, *Test Results*. Cranfield Institute of Technology, Cranfield, England.

Cleverdon, C.W. (1977). *A Comparative Evaluation Of Searching by Controlled Language and Natural Language in an Experimental NASA Data Base*. Report ESA 1/432, European Space Agency, Frascati, Italy.

Dabney, D.P. (1986). The curse of Thamus: An analysis of full-text legal document retrieval. *Law Library Journal, 78(5)*, 5–40.

Flores, F., Graves, M., Hartfield, B. and Winograd, T. (1988). Computer systems and the design of organizational interaction, *ACM Transactions on Office Information systems, 6(2)*, 153–172.

Lancaster, F.W. (1968). *Evaluation of the Medlars Demand Search Service*. National Library of Medicine, Bethesda, MD.

Langendoen, D. Terence and Postal, P. (1984). *The Vastness of Natural Languages*, Basil Blackwell, Oxford, U.K.

Malone, T., Grant, K.R., Turbak, F.A., Brobst, S.A. and Cohen, M. (1987). Intelligent Information-Sharing Systems. *Communications of the ACM, 30(5)*, 390–402.

Maron, M.E. (1988). Probabilistic design principles for conventional and full-text retrieval systems. *Information Processing and Management, 24(3)*, 249–256.

Nelson, M.J. (1988). Correlation of term usage and term indexing frequencies. *Information Processing and Management, 24(5)*, 541–548.

Salton, G. (1970). Automatic text analysis. *Science*, 168, 3929, 335–343.

Salton, G. (1973). Recent studies in automatic text analysis and document retrieval. *Journal of the ACM, 20(2)*, 258–278.

Salton, G. (1986). Another look at automatic text-retrieval systems. *Communications of the ACM. 29(7)*, 648–656.

Sparck Jones, K. (1981). *Information Retrieval Experiment*, London: Butterworths.

Stanfill, C. and Brewster, K. (1986). Parallel free-text search on the connection machine system, *Communications of the ACM, 29(1)*, 1229–1239.

Swanson, D.R. (1960). Searching natural language text by computer. *Science*, 132, 3434, 1099–1104.

Swanson, D.R. (1977). Information retrieval as a trial and error process. *Library Quarterly, 47(2)*, 128–148.

Van Rijsbergen, C.J. (1979). *Information Retrieval*. (2nd ed.) London: Butterworths.

Verity Inc. (1988). Corporate Backgrounder. Mountain View, CA.

Wall, E. (1964). Further implications of the distribution of index term usage. *Proceedings of the American Documentation Institute, 1*, 457–466.

Winograd, T. and Flores, F. (1987). *Understanding Computers and Cognition: A New Foundation for Design*, Reading, MA: Addison-Wesley.

Zipf, G.K. (1949). *Human Behavior and the Principle of Least Effort*. Cambridge, MA: Addison-Wesley.

Zoellick, W. (1986). Selecting an approach to document retrieval. S. Ropiquet, Ed., *CD ROM: Optical Publishing*, (Chpt. 5). Redmond, Washington: Microsoft Press.