STAIRS Redux: Thoughts on the STAIRS Evaluation, Ten Years after

David C. Blair
Computer and Information Systems Department, Graduate School of Business, The University of Michigan, Ann Arbor, MI 48109-1234. E-mail: dcblair@umich.edu

The test of retrieval effectiveness performed on IBM's STAIRS and reported in *Communications of the ACM* ten years ago, continues to be cited frequently in the information retrieval literature. The reasons for the study's continuing pertinence to today's research are discussed, and the political, legal, and commercial aspects of the study are presented. In addition, the method of calculating recall that was used in the STAIRS study is discussed in some detail, especially how it reduces the five major types of uncertainty in recall estimations. It is also suggested that this method of recall estimation may serve as the basis for recall estimations that might be truly comparable between systems.

Introduction

The results of the evaluation of IBM's STAIRS (*ST*orage *A*nd *I*nformation *R*etrieval *S*ystem), a full-text document retrieval system, were published ten years ago (the principal findings of the study were reported in Blair and Maron [1985]; additional results and a deeper discussion of the evaluation methodology were presented in Blair [1990]). In the rapidly changing field of computerized information technology, system tests a decade old are ancient history, and it is rare that empirical research even a few years old maintains relevance to current work. Yet the STAIRS evaluation made an impact that, judging by its continuing high citation rate, remains relevant to today's work in the field of Information Retrieval. Just last Spring one of the authors gave an extensive interview about the STAIRS study for the lead article of *Law Office Computing* (Bauman, 1994), and in a 1993 article in *JASIS*, Paijmans stated, "Cleverdon's observations on the subject of human indexing and the famous Blair/Maron experiment in full text retrieval still loom darkly over all attempts to substantially alter the effectiveness of information retrieval techniques" (Paijmans, 1993). There are a number of reasons for this continuing interest in the STAIRS study, and not all are as foreboding as Paijmans' observation. As Gey and Dabney pointed out, "The Blair and Maron study stands alone as the only large-collection study whose validity is not questioned" (Gey and Dabney, 1990).

The STAIRS evaluation was a rare attempt to benchmark a comparatively large, commercial information retrieval system. Exhaustive recall studies of even small information retrieval systems are notoriously capital and labor-intensive, and an exhaustive study of a large, operational system is beyond the budget of most institutions—in current dollars, the STAIRS study would probably cost more than $500,000. But the organization for which the STAIRS evaluation was carried out had not only an interest in the outcome of such an evaluation, it also had deep enough pockets to fund such a study. More importantly, it saw that the success or failure of large corporate lawsuits was often dependent on how well the information germane to those lawsuits was managed. Due to the large size of the company and the large number of contracts it was a party to, the company was regularly engaged in litigation. (It kept its own corporate counsel within the company, as well as a major San Francisco law firm on retainer.) These lawsuits were often of substantial size, so finding an effective litigation support system, such as STAIRS, was of paramount importance, not just for the lawsuit that we studied, but for all pending and future lawsuits. The STAIRS study was an attempt to evaluate an information retrieval system used in the defense of a $237,000,000 lawsuit, so the cost of the study, while high compared to most information retrieval evaluations, was considered a reasonable expense compared to the overall risk of the lawsuit.

Another reason for the continuing interest in the STAIRS study is that the majority of today's commercial information retrieval systems employ simple full-text retrieval techniques that are very similar to those used by STAIRS. (These simple full-text techniques date back to the 1950s in experimental systems.) To understand why

 CCC 0002-8231/96/010004-19

this is, we must distinguish between "intellectual" and "physical" access to documents. The methods of intellectual access, of which there are many, are used to select or rank available documents based on matching the searcher's request with the representation of the documents' intellectual content. The techniques of physical access, on the other hand, are concerned with how to deliver the selected documents to the searcher. Physical access deals with the complex issues of formatting documents, managing their transmission across networks, combining documents from different systems, etc. Obviously, intellectual access is a prerequisite for physical access; this is why information retrieval research focuses primarily on improving the techniques of intellectual access to documents. The specific problems of physical access are more of an engineering problem and are only rarely addressed within the information retrieval literature [Salton, 1989; Van Rijsbergen, 1979]. But in spite of the central importance of intellectual access, commercial document retrieval developers have applied their resources much more vigorously to the problems of physical access than to the problems of intellectual access. The primary reason for this is that improvements in physical access, even on large commercial systems, are comparitively easy to measure. Advances in intellectual access are much more difficult and costly to estimate—as the STAIRS study showed. A harsh reality of commercial investment is that venture capital flows towards success. Since physical access improvements can be measured and compared fairly precisely, while advances in intellectual access cannot, it is easy to see that most investment capital will back the measurable successes of physical access, rather than the much less quantifiable successes of intellectual access.

The question then becomes, with the widespread knowledge of the STAIRS evaluation why do so many commercial document retrieval systems use simple full-text retrieval? In the first place, there are few evaluations of alternative document retrieval techniques on large-scale systems that could be compared with the STAIRS study, so there are no obvious, proven replacements for simple full-text retrieval. Further, there were some researchers even in the information retrieval community who claimed that the low recall rates found in the STAIRS study are typical of all retrieval systems: ". . .not only is [the STAIRS evaluation's] level of performance typical of what is achievable in existing, operational retrieval environments, but that it actually represents a *high-order* of retrieval effectiveness" (Salton, 1986).

Finally, because of the widespread use of word processing, most documents now *begin* in computer-readable form. As a result, simple full-text retrieval techniques can be implemented very easily, becoming the method of choice by default.

The difficulty in measuring advances in intellectual access has a further consequence, namely, that without a clear measurement of success, not only do we not see which systems are better at intellectual access and which are worse, it is also difficult for information retrieval research to build on past successes in this kind of access and avoid past mistakes.

It has been estimated that the commercial text retrieval industry was worth $232 million in 1992 and has continued to grow at a 35% annual rate making their estimate of the industry's worth in 1995 to be $570 million—with the same rate of growth continuing in the near future (Delphi Consulting Group, Inc., 1992). Of course, there *have* been breakthroughs in the storage and transmission of documents—techniques of physical access—but these advances do not improve the intellectual access to documents in any major way (Blair, 1984b). As a consequence, the buyers of information retrieval systems often do not see any compelling reason to invest in advanced information retrieval technology—often called "concept retrieval," as contrasted with simple full-text retrieval, or keyword systems with assigned document descriptions. Simple full-text retrieval, or retrieval based on assigned or automatically generated keywords requires some kind of a match between the terms in the search query and the terms used to represent the documents. (In full-text systems, the words which represent a document are the words in the document that discuss the intellectual content of that document.) "Concept retrieval" is based on statistical (e.g., term co-occurrence) or semantic associations between the various terms used to represent the documents. These associations make it possible for retrieval systems to return documents, or document representations, which do not contain any of the terms specified in the search request, but which contain terms statistically or semantically related to those terms. In the vernacular of retrieval theory, these methods can retrieve documents which match the intellectual "concept" implied by the search terms, even though none of the search terms is used to represent the retrieved documents. Of the 85 commercial information retrieval products listed in Delphi's industry survey, only five are listed as having "concept based retrieval"—systems such as Information Access Systems' ITMS (*I*ntelligent *T*ext *M*anagement *S*ystems), Thunderstone's Metamorph, and Verity's TOPIC (Delphi Consulting Group, Inc., 1992). A majority of commercial information retrieval systems—including the market leader, IDI's ZyIndex—remain simple, full-text systems that, from the point of view of intellectual access, are basically the same as STAIRS.

The valuable TREC conferences, which have sponsored several comparisons of information retrieval techniques, have shown that most of the advanced "concept"-oriented research systems perform at a similar, modest level of intellectual access. This may not be enough of an improvement over simple full-text systems to justify the additional investment that such a system would require. Researchers such as Sembok and van

Rijsbergen have flatly declared that the statistical techniques on which most advanced systems are based have reached the point of diminishing returns: "The keywords approach with statistical techniques has reached its theoretical limit and further attempts for improvement are considered a waste of time" (Sembok and van Rijsbergen, 1990). So commercial information retrieval system developers find themselves caught between two models: The simple full-text retrieval system, which, by default, is the foundation for many commercial document management systems, and the more advanced "concept"-based systems that are having a hard time convincing customers and investors that their apparently modest improvement over earlier systems is worth the additional investment.

The following discussion will focus on two broad areas: Firstly, we will consider some issues specific to the STAIRS evaluation which have not been discussed before; and, secondly, we will examine some further issues of information retrieval evaluation within the broader context of commercial applications.

The STAIRS Study

While the STAIRS evaluation was originally presented primarily as an objective measurement of Recall and Precision, the retrieval system that we tested had a much larger context—there were also political, legal, and commercial dimensions to the study. These political, legal, and commercial forces had vested interests in the success of STAIRS as a litigation support tool.

Political Aspects

The political dimension of the STAIRS study arose from the nature of the lawsuit. The corporation using STAIRS was a defendant in a lawsuit between the City of San Francisco and the consortium of engineering contractors charged with building the Bay Area Rapid Transit (BART) system. The plaintiff, the City of San Francisco, charged that the dramatic cost overruns, the repeated failures to meet construction deadlines, and the inability to satisfy many of the performance objectives for BART were violations of the contract for building it. At the time of the lawsuit (the late 1970s) the pros and cons of the BART construction were being actively debated in the San Francisco newspapers, and there were frequent incriminations about who, in city government, was responsible for failing to monitor the contractors' performance during construction. As a result, the lawsuit using STAIRS, as well as other pending law suits, contained many documents that were enormously sensitive, politically. Further, the defendant had reasons to successfully defend this lawsuit that went beyond the lawsuit itself. The loss of this case might seriously affect the organization's credibility on other major contracts that it was negotiating at that time.

In most lawsuits, a preferred strategy is to have the other side settle out of court in your favor during the pre-trial phase of the suit. This generally means that you can get the outcome that you want without paying court costs or enduring the appeals that may result from the decision in the first trial. To get the other side to settle out of court, you must convince them that they would be likely to lose their case were it to go to trial—you must convince them, during the pre-trial period, that the law and the evidence will not be likely to uphold their position. But, as corporate lawsuits got larger, another strategy for settling out of court arose. This was to convince the other side in the suit that you could manage the large amount of information germane to the lawsuit more effectively—the implication being that good information access to intellectual content was a necessary condition for effectively defending or prosecuting the case. This superiority in intellectual access could be demonstrated during the pre-trial discovery process. As one of the lawyers involved in the suit we worked on surmised, corporate lawsuits had grown so large that they were then being won or lost not so much on the legal issues involved, but on how effective the information access was, that is, how effectively could the opposing sides find information to defend or prosecute the case. If there was a large difference between the information access capabilities of the opposing sides in a large lawsuit, then the side with the poorer information access capability might perceive its legal position as proportionally weaker. This is not too surprising since the ability to defend or prosecute a lawsuit is contingent on the ability to find evidence to support your position. Increasingly, in corporate lawsuits, this evidence is textual in nature—hence the importance of document retrieval. This is exactly what transpired in the lawsuit we were working on. The company that we were working for was able to convince the plaintiff that their, the defense's, control of information germane to the suit was superior to the plaintiff's. This was one of the factors in the plaintiff's decision to settle out of court. Since the lawsuit we worked on had just been settled by the time we conducted our study, our findings did not affect the outcome of this case. But our company was involved in several other major lawsuits with sensitive political overtones, so the perception of STAIRS' effectiveness for litigation support was a very important, if not crucial, factor in this drama. At the time, some of IBM's lawyers stated that one of the reasons for IBM's recent victory in a large and lengthy lawsuit was the use of STAIRS to manage the information for IBM's side of the lawsuit. It was clear that one of the market niches that IBM had targeted for STAIRS was large scale litigation support. Within this heated atmosphere, it was readily apparent that an objective study showing that STAIRS *did not* provide good intellectual access to its document collection could undermine the defense's ability to force out-of-court settlements in the pending lawsuits. Further, it could also undermine STAIRS' reputation as a

tool for litigation support. Not only might the company we worked for lose pending lawsuits and suffer substantial financial penalties, but they also might have to endure the scrutiny that would occur as politically sensitive documents concerning public construction projects came into public view when the plaintiff made its case. As a result, there were enormous pressures to demonstrate that STAIRS was an effective tool for intellectual access to documents of critical importance.

Legal Aspects

Corporate law firms, faced with the dramatic increase in the size of their typical lawsuit, were looking for a "magic bullet"—a retrieval system that provided effective access to any information germane to a lawsuit. STAIRS appeared to be an ideal system. It not only had enough capacity to provide access to millions of documents, if required, but because STAIRS was a simple full-text retrieval system there was no front-end indexing cost for the lawyers. It looked like the STAIRS user could get state-of-the-art retrieval capability without having to spend the time and effort working out and implementing an indexing structure that would provide access to the intellectual content of the stored documents. The prevalent attitude of document retrieval users was the conviction that the full-text of a document captured the "complete meaning" of the document. Nothing more would be needed for effective retrieval. In short, it looked like STAIRS might give you something for nothing.

Commercial Aspects

STAIRS was, first and foremost, a commercial product of the IBM Corporation. With document management and retrieval becoming a major concern of business, IBM was well positioned to capture a large percentage of the mainframe document retrieval market. STAIRS was a relatively new system, and had not yet attained the public confidence that other IBM systems, like the IMS data base system, enjoyed. Any negative publicity might seriously undermine the growing, positive reputation of STAIRS, especially in its use for large-scale litigation support.

The Ethics of Retrieval Evaluation

Traditionally, information retrieval tests have been seen from a purely scientific perspective. The tests of retrieval effectiveness are simply attempts to calibrate a technical process. True, relevance judgments may add a subjective component to this evaluation, but, in general, the only concern of the evaluators is how to measure the performance of the system in question. The dictates of the scientific method demand that researchers rigorously uphold their scientific objectivity—they must be concerned primarily with the accuracy of their measurements, and not whether the system performs poorly or well. It can be the case that the evaluators may *prefer* one outcome over another, but those preferences should not influence or qualify the outcome of the evaluation. But as information retrieval systems are utilized more frequently in commercial situations, they begin to take on the values and interests of the applications in which they are involved. This may call into question the objectivity of an evaluation.

Up until the STAIRS evaluation, the types of retrieval systems that were tested were primarily small experimental systems, or library/bibliographic systems. Systems like these were comfortable with the scientific objectivity of any evaluations, in part because these systems were often closely associated with academic or research institutions. It was also the case that the results of these evaluations did not affect the viability of the system—that is, no library-based retrieval system was shut down or lost money because it was found to have poor searching capability. (Ironically, poor searching methods can be a financial boon for commercial information providers if they bill their clients based on connect time, which many do. That is, the poorer the retrieval method, the longer the client must spend searching for what he wants, and the longer he searches, the more he has to pay for that service.)

Recent advances in commercial information retrieval systems have taken information retrieval outside of the traditionally supportive academic or research environment. The survival of many of these commercial systems may be dependent on the outcome of any evaluation of retrieval effectiveness. A test showing that a system does not perform well might be reason enough to shut it down (a utility company was forced to shut down four nuclear reactors—at a cost of $2 million per day in lost revenue—because its document retrieval system could not retrieve important safety manuals and schematics quickly and reliably [Fleischer, 1990]). In the STAIRS study it was clear that if our evaluation showed that the system was performing poorly, and that such a low level of performance could not be easily remedied, then both the organization that owned the system and the vendor that sold it (IBM) might incur substantial penalties. If the results were made public, the defendant might have a harder time settling pending lawsuits out of court, and IBM would incur a lot of bad publicity for STAIRS at a time when they were trying to build support for this relatively new system. Clearly, a positive result for the STAIRS evaluation would have made everyone involved with STAIRS and the lawsuit a lot happier: IBM would get the good publicity that it desired, the defendant could be assured that STAIRS was a good litigation support system, and we, as evaluators, would have a recall study confirming the advances claimed for a major software product. (IBM claimed that internal studies showed that STAIRS could achieve recall values in the 80–90%

range.) Unfortunately, our study did not confirm IBM's optimism.

As more information retrieval studies focus on commercial applications, researchers must be aware of the commercial pressures to which they may be subjected. In fact, these pressures may put the researcher in an ethical dilemma. For example, suppose that a researcher is employed by a software vendor to test its product—an information retrieval system. As an employee of the software company, the researcher has a personal interest in having his/her test produce good results. He or she will have done their job, and both the researcher and the vendor will profit from the good publicity that would follow. But what if the researcher's evaluation finds that the information retrieval system performs poorly, and such poor performance cannot be easily improved? If those results are made public, sales may fall off and the researcher could conceivably lose his/her job. If the poor results of the study are suppressed, though, the researcher, knowing that the unpublished results reflect badly on the system, may be pressured to dissemble or remain silent, and the clients will not be told the true performance levels of the system. One can make this scenario even more compelling if the information retrieval system in question is used in a critical area of retrieval where poor retrieval results might cause significant penalties—such as with legal or medical information, or the nuclear power plant scenario described above.

What the information retrieval researcher or developer must remember is that to develop a working commercial information retrieval system requires substantial investment. Up until recently, this was not a factor in retrieval evaluations. As more and more commercial systems are evaluated, the information retrieval system evaluator must understand that there may be enormous pressure by the investors to protect their investment (that is, there will be a lot of incentive to have a "good" evaluation for their system). If you combine these pressures with the inexact techniques of a recall/precision study, it is not hard to see that such an evaluation may become biased, even if the evaluator's intent is to be objective.

The Problematic Nature of Recall Studies

One of the major problems with recall studies is the large amount of subjectivity or uncertainty in such assessments. The nature of the scientific method demands as much objectivity and certainty as possible, but if we look at the history of retrieval evaluation we find that in most studies there remains a substantial amount of uncertainty in even the most carefully conducted evaluation. This uncertainty manifests itself in five principal ways:

1. The relevance judgments of the searcher—what relevance means operationally, and how consistently these judgments can be made by the searchers.
2. The stopping point for the searcher. (The stopping point could result from either getting enough of what he wants, or it may be the point at which he gives up in futility. We might call this the searcher's *degree of persistence.*)
3. Where the evaluator should look for unretrieved, relevant documents. (This implies a rank-ordering of the best places to look.)
4. The relevance judgments used by the evaluator.
5. The stopping point for the evaluator (either, the point where he has found "enough" unretrieved relevant documents, or where he reaches his own futility point). This is an indication of how persistent the evaluator is.

The more uncertainty there is in the experimental design of a recall/precision study, the less decisive or reliable it becomes. For example, if relevance judgments are known to be uncertain, then the estimated recall values must be necessarily imprecise. As a result, the evaluator cannot reject the hypothesis that high recall values may not actually mean that the system is performing well. These high values may have resulted from some or all of the five uncertainties described previously.

For a recall/precision study to be reliable, the variation in recall levels must be due to the retrieval performance of the system rather than any of the uncertainties (above). The necessary, but perhaps not sufficient, condition for a reliable recall study must include the reduction—ideally, elimination—of these uncertainties. This is no easy task, but it was foremost in our minds as we designed our test of STAIRS.

Relevance

The problems with relevance judgments are well known in the information retrieval literature so they will not be repeated in detail here. Basically, there are two problems with relevance:

1. Does relevance measure what is important in a search?
2. Are the relevance judgments that are made consistent?

In considering question 1, we agreed with Cooper (1973) and Swanson (1977) that relevance typically measures "topicality" and was *not* what we wanted to measure. We agreed with them that what we really wanted to measure was the *utility* of the retrieved documents. STAIRS was, first and foremost, being used *to do* something—it was an essential part of a well-defined information-intensive activity, the defense of a lawsuit. It was also clear that accurate judgments of document utility could only be made by those who were involved in the activity being supported, i.e., those who originally submitted the que-

ries. That meant that no matter how difficult or expensive it was to have the lawyers' estimations of document utility/relevance, we needed to have them if we were to have a good estimation of STAIRS ability to return useful documents.

The second question, concerning the consistency of the relevance judgments, was easy to test. We simply sent the lawyers the same retrieved sets of documents several times over the course of the 6 months that it took us to conduct the test. Since the test generated hundreds of retrieved sets of documents, the lawyers' evaluations of these redundant document sets were typically independent. For most of the experiment, the lawyers were receiving hundreds of documents a day to evaluate, in addition to their duties on other lawsuits, so that if they saw a retrieved set of documents twice over a 2-month period, they generally did not remember that they had seen the same set of documents before. (To further insure the objectivity of the experiment we did not resubmit any document sets to the lawyers until after the recall values had been calculated.) What we found was that the lawyers were rigorously consistent in their assessments of document utility. (The lawyers ranked the documents as "vital," "relevant," "marginally relevant," or "not relevant." This was the classification scheme that they felt comfortable with. Even though they described the documents as "relevant," they were well aware that they were really assessing document utility. That is, if a document was retrieved twice it was only relevant/useful the first time it was seen.) There was a slight, but not statistically significant, tendency for documents to move down in category over time (e.g., from "vital" to "relevant"), but it was never the case that a relevant document became non-relevant, or vice versa).

When Should the Lawyers Stop Searching?

We also left the endpoint of the search up to the lawyers. We simply told them to search until they found enough useful documents, in their estimation, to conduct the defense of the lawsuit. Since the lawyers were the ones who would defend the lawsuit, they were naturally the only people to know when to stop. In addition, the lawyers volunteered a quantitative stopping point, feeling that they needed at least 75% of the relevant documents, and 100% of the vital documents to feel confident in the defense of the suit.

How Should the Evaluators Judge Relevance?

The judgment of relevance from the point of view of the evaluators was relatively straightforward. Previously, we had defined relevance as the lawyers' estimation of the utility of retrieved documents. Since this judgment captured the evaluation that we wanted, we simply used the same procedure to judge unretrieved documents as we did for retrieved ones. In other words, *all* judgments of relevance during the test were made by the lawyers who originally formulated the queries.

When Should the Evaluators Stop Searching for Unretrieved Relevant Documents?

While we were able to reduce three of the uncertainties (above, numbers 1, 2, and 4), the determination of where we, as evaluators, should look for unretrieved, relevant documents (number 3), and how long we should persist in our search (number 5) were more problematic. Our initial criterion for stopping was to search until we found a significant number of unretrieved, relevant documents, or, failing that, when we had reached a point of diminishing returns. Fortunately, we were able to find a significant number of unretrieved, relevant documents, so our concern about when to stop turned out to be less of an issue, as far as STAIRS went—that is, as long as we found a significant number of unretrieved, relevant documents, our test results would be informative. But in the backs of our minds, we were still concerned with defining, in our experimental protocol, a general procedure for determining the search space for unretrieved, relevant documents and a systematic method for searching that space efficiently. Even though we found many unretrieved, relevant documents, there were some queries where it was not easy to do this, so we wanted to consider, first, only those sets of unretrieved documents which were likely to contain significant numbers of relevant documents. This would make our searches more efficient and would give us a rationale for stopping our search if the number of unretrieved, relevant documents fell off significantly.

The literature on recall studies contains many discussions about the first, second, and fourth uncertainties (above), though more on the first than the others. But there is almost no discussion of the third and fifth types of uncertainty. In our estimation, these are the most important uncertainties in recall evaluations, and are possibly the source of the greatest variation in recall studies. That is, while it is possible to compare and standardize the relevance judgments of the searchers and evaluators (e.g., that the documents should be useful in some well-defined task), and define an endpoint for searching (e.g., that one should search until he finds enough documents to enable him to complete his task, or, failing that, until he reaches a futility point [Blair, 1980]), there have been no attempts to define, formally, a search space or an endpoint in the search for unretrieved relevant documents. There is also almost no discussion in the information retrieval literature about a general methodology for finding relevant, unretrieved documents. Of course, since many previous recall studies have been conducted on systems with less than a thousand documents, the size of the system has not been a major factor (the evaluators could search the entire collection if need be). But, increasingly, information retrieval techniques are being used to man-

age large collections of documents—collections that are not only too big to search entirely, but are even too big to sample from with high confidence levels (Tague, 1981). The large size of such collections requires that we develop a systematic strategy for finding unretrieved, relevant documents, otherwise our searches would be largely ad hoc and we could not reject the hypothesis that we had entirely missed a substantial number of unretrieved, relevant documents. Another reason to have a systematic procedure for either finding unretrieved, relevant documents or knowing when to stop looking when none is found, is that such a procedure would make recall estimations on different systems comparable. Currently, there is so much unexplained variance in recall evaluations, that it is unclear whether any recall estimations are comparable at all. The variation in both the persistence of the evaluators and where they looked in their search for unretrieved, relevant documents are probably the most important factors in the differing results of recall studies on large documents retrieval systems.

The general interpretation of recall is, of course, that a high recall value is taken to mean that there are few unretrieved, relevant documents in the collection. But such high recall values can also result from a weak effort in trying to find unretrieved, relevant documents. In some cases, high recall values may even *reward* a poor effort on the part of the evaluator. That is, a sloppy, naive, or cursory attempt to find unretrieved, relevant documents will inevitably result in higher recall values than a more persistent effort. On the other hand, exceptional persistence in the search for unretrieved, relevant documents will usually turn up more relevant documents, which, in turn, will drive the value of recall down. Since information retrieval is generally assumed to be easier than it actually is, any studies that put recall at a realistic level will probably be seen as bad news. In short, an exceptionally good search effort may result in low recall values, while a weak, inexperienced, or naive search effort may be rewarded by high recall values. A sloppy evaluator can make an entire career out of bad recall estimations, since he is continually the bearer of good news.

The lack of persistence by the evaluator has at least three causes. In the first place, the evaluator may believe that document retrieval is a relatively straightforward process. The root of this presumption is a failure to distinguish the data retrieval model from the document retrieval model (Blair, 1984a, 1990). Data retrieval is characterized by relatively precise queries and relatively precise descriptions of the stored data (Buck Mulligan's address is simply "Buck Mulligan's address," there is little ambiguity in the representation of such data). So data retrieval is usually a process of simple matching. Document retrieval, on the other hand, is a much less precise process, especially when the search is for documents with a certain intellectual content (what is the precise query that will retrieve documents that discuss Central European investment prospects? It is unlikely that there is a single, precise, content-oriented query to do this). The evaluator who sees document retrieval as a form of data retrieval will assume that document retrieval is more straightforward than it actually is and that difficulties in searching are exceptions to the rule. If initial searches for unretrieved, relevant documents are unsuccessful, then the naive evaluator will likely draw the hasty conclusion that there are few unretrieved, relevant documents and not draw the more reasonable conclusion that both his queries and the document representations may be imprecise and difficult to match.

The second reason why an evaluator might not be persistent in looking for unretrieved, relevant documents is that he may have a mistaken goal for his recall study—he may believe that the goal of retrieval evaluation is to *confirm* the effectiveness of the system rather than to make a concerted effort to find evidence *against* its effectiveness. This bias will lead the evaluator to place more confidence in successful queries than in unsuccessful ones. Such a bias originates from a mistaken view of the scientific method, namely, that science seeks *confirmations* of hypotheses rather than *refutations* (here, the search queries submitted by the searcher are taken as hypotheses about how relevant documents are represented. It is supported or refuted by the number and quality of relevant documents it returns [Blair, 1982]). But, as Popper [1959, 1968] pointed out, confirmations of hypotheses are easy to get and do not tell us very much unless there is some risk involved in their prediction. The key to the scientific method is to be persistent in the attempt to *refute* the hypotheses in question, and, having failed to refute them, the investigator can accept them provisionally. Swanson observed some time ago that on a large information retrieval system, it is relatively easy to find documents that satisfy the user's query. In fact, it is hard to write reasonable queries that *do not* retrieve at least some relevant documents. This he called the "fallacy of abundance." He called it a "fallacy" because it leads the searcher to the comfortable conclusion that a system works well:

> A scientist who nowadays imagines either that he is keeping up with his field or that he can later find in the library whatever may have escaped his notice when it was first written is a victim of what might be called the "fallacy of abundance." The fact that so much can be found on any subject creates an illusion that little remains hidden. Although library searches probably seem more often than not to be successful simply because a relatively satisfying amount of material is exhumed, such success may be illusory, since the requester cannot assess the quantity and value of relevant information which he fails to discover. (Swanson, 1960)

This is precisely the position the lawyers in the STAIRS study found themselves in. The relative ease by which they retrieved *some* relevant documents gave them the

false impression that they had gotten all, or nearly all, of the documents that satisfied their need. The goal of retrieval according to Swanson is not just to retrieve *any* relevant/useful documents, but to find the *most useful* documents first.

The final reason why evaluators may not be as persistent as they need to be is that they do not look in the best places for unretrieved, relevant documents. That is, they just do not know where to start. After a few naive attempts to find unretrieved relevant documents turn up empty, the evaluator runs out of ideas about where to look and gives up, thereby insuring a high recall estimation.

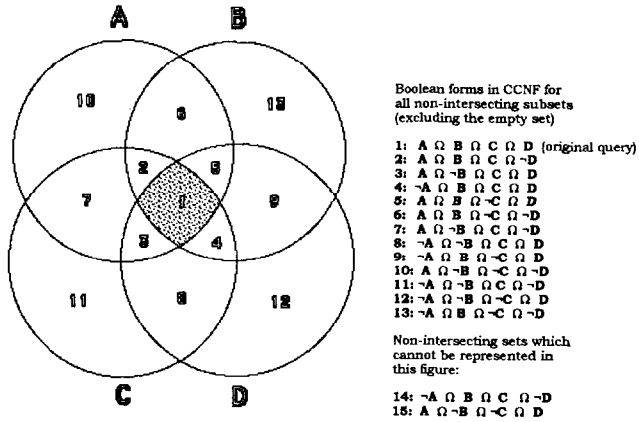

FIG 1. All candidate sets in CCNF.

Where to Look and when to Stop Looking: The Roles of Logical Modification and Semantic Expansion

In an attempt to formulate a systematic procedure for finding unretrieved, relevant documents, we developed a logical method for systematically delineating subsets of the document collection that might be "rich" in unretrieved, relevant documents. (That is, if there *are* unretrieved, relevant documents, then they are very—perhaps, most—likely to be in these subsets.) Since the document collection under STAIRS' control was reasonably large (the text for about 350,000 pages of information on-line), many of the queries that the lawyers used were conjunctions of words or phrases. (Most entirely disjunctive queries retrieved too many documents to be useful. Even single-term queries retrieved large numbers of documents (over 10,000 documents for one keyword)—that is, these retrieved sets were larger than the searcher's "futility point" [Blair, 1980], and, thus, could not be examined.) Let's suppose the set of documents that the lawyers were happy with was retrieved with the following query (in actuality, the lawyers submitted a number of queries to retrieve the documents that they wanted for any single search, but for simplicity, we will assume that only one query was satisfactory):

A Ω B Ω C Ω D

(Where Ω is the operator for conjunction and A, B, C, & D are either keywords or disjunctions of keywords.) This is what is known as "Conjunctive Normal Form" (CNF) in propositional logic:

> A search query is in CNF when, in addition to the symbols representing the search terms (A, B, . . .) it contains no other symbols than those for conjunction (Ω), disjunction (v) and negation (¬). The negation symbol can only apply to single terms, and the conjunctions are applied to either single terms, negated single terms, or disjunctions of single terms. [Copi, 1965]

The following are examples of CNF:

A
A Ω B
A Ω ¬B
A Ω (B v C) Ω D

All statements (i.e., search queries) in Propositional or Boolean logic can be transformed into CNF.

Since all the queries submitted to STAIRS were formulated in propositional (or, Boolean) logic, they were all convertible into conjunctive normal form, by definition (see Blair, 1988 for a brief discussion of conjunctive normal form in retrieval queries). But the above query is not just the query that retrieved documents that the lawyers wanted, it is also the basis for describing unretrieved sets of documents that have a high probability of containing a significant number of relevant documents. While this query defined the retrieved set of documents, candidates for unretrieved sets of documents (hereafter called "candidate sets") were:

A Ω B Ω C Ω ¬D
A Ω B Ω ¬C Ω D
A Ω ¬B Ω C Ω D
¬A Ω B Ω C Ω D

Including the negation ("¬") of some of the terms in the search query instead of leaving the terms out produces a form of representation known as "Complete Conjunctive Normal Form" (CCNF). These complete conjunctive normal form representations describe non-intersecting sets of documents that are semantically close to the search query that generated the retrieved set, but do not contain any of the documents in the original retrieved set. (See Fig. 1). Because of the inherent indeterminacy of keywords, both as search queries and as document representations, the original conjunction that returned the retrieved set is semantically imprecise (Blair, 1986). This imprecision means that the sets formed by taking

the successive negation of one or more terms in the original query are good candidates to contain relevant, unretrieved documents. In effect, the indeterminacy of keywords can be exploited by the evaluator. Similar candidate sets of unretrieved documents can be formed by successively negating *two* of the keywords in the original query, or, further, by successively negating *three* of the keywords in the original query. (By negating two keywords, six additional non-intersecting candidate sets can be created, while negating any three keywords can create four additional candidate sets for a four-term CCNF query. This gives the evaluator a total of 14 non-intersecting sets of unretrieved, possibly relevant documents based on the original four-term query.) For example, negating two of the terms in the original four-term query yields the following six non-intersecting sets:

A Ω B Ω ¬C Ω ¬D
A Ω ¬B Ω C Ω ¬D
¬A Ω B Ω C Ω ¬D
A Ω ¬B Ω ¬C Ω D
¬A Ω B Ω ¬C Ω D
¬A Ω ¬B Ω C Ω D

As a general rule, the number of candidate sets that can be derived from n terms is 2^n–2. (More formally, this is the power set minus the empty set [where all terms are negated] and minus the set in which no terms are negated [the original query]). It may be the case that on a large retrieval system these candidate sets have relatively large numbers of documents in them, which would preclude examining them in their entirety. This was the case in the STAIRS study. To deal with this, we simply sampled from these candidate sets of documents to get an estimate of how many relevant documents existed in them. This, of course, gave us a *maximum* value for recall since we were sampling from subsets of the document collection, rather than all of it. It was also the case in the STAIRS study that individual searches were composed of several distinct queries. In this case, we simply performed the logical modification on each of the distinct queries and sampled each of the resulting candidate sets. The estimation of recall, then, was the union of the estimated unretrieved relevant documents for all of the candidate sets. While the percentage of unretrieved, relevant documents in the entire document collection is too small to sample with confidence (Tague, 1981), the candidate sets that we have created here will often be small enough and rich enough in unretrieved, relevant documents to be able to sample at a high confidence level. If they do not have a high concentration of unretrieved, relevant documents in any of these logically modified candidate sets, then it is unlikely that there will be many more relevant documents than were in the original retrieved set. There are exceptions to this observation, of course, such as when the original search query is so naive or ineffective that it misses virtually all of the desired documents. But we found in the STAIRS study that the majority of unretrieved, relevant documents were to be found in these logically derived candidate sets. (N.B., the diagram in Fig. 1 does not include all the candidate sets. Sets for

A Ω ¬B Ω ¬C Ω D
¬A Ω B Ω C Ω ¬D

are not represented due to the limitations of two-dimensional Venn diagrams.)

The Importance of Complete Conjunctive Normal Form (CCNF)

By generating non-intersecting sets of documents through CCNF, the evaluator attains a number of subtle, but important, advantages. First of all, as she searches through these candidate sets for unretrieved, relevant documents, she will not see documents that she has already seen before. This makes her search much more efficient. The generation of non-intersecting candidate sets is accomplished by including the negation of the keywords in queries which describe them. Without these negations, the more general queries (e.g., A Ω B) will always contain the documents retrieved by the more restrictive queries (e.g., A Ω B Ω C), so the evaluators are constantly seeing documents that they have seen before. This makes the search for unretrieved, relevant documents very inefficient and prone to error. By not using the CCNF versions of the queries to generate the candidate sets, the evaluators must keep track of which documents have been seen and which have not, at each stage in the evaluation process for that query. The CCNF versions of the evaluators' queries give them non-intersecting sets of documents automatically, greatly simplifying the search effort. It also gives the search for unretrieved relevant documents a well-defined, finite search space that, in turn, provides evaluators with clear, non-redundant candidate sets to look through for unretrieved, relevant documents. Finally, by excluding previously retrieved documents, the candidate sets will be smaller. This makes it less likely that the evaluator will reach his/her futility point.

Of course, it is not necessary for the evaluator to search through or sample from all candidate sets of queries. Lack of time and/or money may preclude an exhaustive analysis. In such a situation, the evaluator may specify his/her degree of persistence by limiting his/her examination to the candidate sets of queries defined by the negation of a single element (in our example, this would yield four candidate sets). We might call this a "Level One" recall estimation. The advantage of distinguishing different levels of evaluation is that it might permit comparisons between recall estimations of different degrees of persistence—something we have not had until now. That is, a "Level One" recall estimation could still be compared with a more comprehensive recall estima-

tion of, say, "Level Three," because the Level Three estimation is really a combination of Levels One, Two, and Three.

An Objection

One obvious objection to this method of creating candidate sets is that it is complicated to do and, when all the candidate sets are searched, it is no different from searching or sampling from the disjunction of the original query terms: AvBvCvD. Why not save time and just sample from this disjunctive version of the query? On a small document retrieval system this may be possible because the disjunctive version of the query may be small enough to sample from its entirety. But we are concerned here with large scale systems, systems in which the disjunction of all, or even a few, of the terms may be too large to even sample from (remember our comment that one term in the STAIRS system we studied retrieved 10,000 documents). The use of CCNF provides a series of candidate sets that give the evaluator more options in his/her recall estimation. The evaluator can choose to look at *all* or just a few of the candidate sets, examining or sampling from only those sets of unretrieved documents that are small enough and rich enough in relevant documents to warrant an examination. Even sets that are too large or not rich enough in unretrieved relevant documents to be examined in their entirety, can be broken down into a series of sets that, in part, may allow such an examination. In our example (Fig. 1), the set "A" is composed of the following candidate sets:

$$A \,\Omega\, B \,\Omega\, C \,\Omega\, \neg D$$
$$A \,\Omega\, B \,\Omega\, \neg C \,\Omega\, D$$
$$A \,\Omega\, \neg B \,\Omega\, C \,\Omega\, D$$
$$A \,\Omega\, B \,\Omega\, \neg C \,\Omega\, \neg D$$
$$A \,\Omega\, \neg B \,\Omega\, C \,\Omega\, \neg D$$
$$A \,\Omega\, \neg B \,\Omega\, \neg C \,\Omega\, D$$
$$A \,\Omega\, \neg B \,\Omega\, \neg C \,\Omega\, \neg D$$

While "A" itself may be too large to even sample from, the above CCNF candidate sets cover the "A" region entirely. Some of these smaller sets may be small enough or rich enough in relevant documents to make searching viable. Of course, in the final analysis, what makes this objection moot is that the reason why the searcher uses a conjunction of four terms in the original query is most likely because any single term or disjunction of terms retrieves far too many documents (Blair, 1980). Such large retrieved sets are likely to be too large for the evaluator to search, also, and perhaps too large to sample from either.

Semantic Expansion

"Semantic expansion" was a variation in our logical modification method made by adding, disjunctively, keywords that were synonymous with the keywords in the original query. For example, given our original keywords (A, B, C, & D) suppose that we found synonymous terms E, F, G, and H, such that E was a synonym for A, F, and G were synonyms for B, and H was a synonym for D. We could add these to our original complete conjunctive normal form variations as follows:

$$(A \vee E) \,\Omega\, (B \vee F \vee G) \,\Omega\, C \,\Omega\, (D \vee H)$$

This yields an expression that is still in CNF since CNF is defined as a conjunction of single terms or a conjunction of disjunctively related single terms. These expressions could be modified to create expressions for the candidate sets in the same way that the original expression without the disjunctions was modified, namely:

$$(A \vee E) \,\Omega\, (B \vee F \vee G) \,\Omega\, C \,\Omega\, \neg (D \vee H)$$
$$(A \vee E) \,\Omega\, (B \vee F \vee G) \,\Omega\, \neg C \,\Omega\, (D \vee H)$$
$$(A \vee E) \,\Omega\, \neg (B \vee F \vee G) \,\Omega\, C \,\Omega\, (D \vee H)$$
$$\vdots$$
$$\vdots$$
$$\text{etc.}$$

Strictly speaking, these expressions are *not* all in CCNF (because the negations in expressions one and three are applied to expressions that consist of more than one term). This is easy to remedy, though, by applying DeMorgan's Theorem. Expressions one and three (above) are converted into the following CCNF expressions, respectively:

$$(A \vee E) \,\Omega\, (B \vee F \vee G) \,\Omega\, C \,\Omega\, \neg D \,\Omega\, \neg H$$
$$(A \vee E) \,\Omega\, \neg B \,\Omega\, \neg F \,\Omega\, \neg G \,\Omega\, C \,\Omega\, (D \vee H)$$

The principal advantage of the logical modification and semantic expansion methods is that they offer us a *systematic* way of retrieving relatively small candidate sets of documents that are likely to contain unretrieved, relevant documents but contain neither the originally retrieved documents nor documents from the other candidate sets. Having a systematic way to delineate candidate sets of documents was an enormous advantage in the STAIRS study. Because the document collection was so large—though not large by today's standards—we could not search it in its entirety for unretrieved, relevant documents. Nor could we sample the entire collection with confidence, as we have already pointed out. As a result, it was imperative that we find a reliable, systematic way to determine likely places to find unretrieved, relevant documents. In a system with a large document collection we could not search in a random or ad hoc manner since there were simply too many possible places where unretrieved relevant documents might be. (As a control, we sampled at random from the document collection at

large, but this method produced no unretrieved, relevant documents.)

A systematic method for finding unretrieved, relevant documents has several advantages: In the first place, it allows the evaluators to search for unretrieved, relevant documents more efficiently than ad hoc methods would permit. It also significantly narrows the search space for unretrieved, relevant documents. (Most of the individual searches conducted by the lawyers in the STAIRS study had somewhere around 100–200 relevant documents in total (retrieved documents plus unretrieved, relevant documents). This meant that only 100/40,000–200/40,000 or 0.25–0.50% of the collection might be relevant to a given query—the proverbial needle in a haystack. To find that "needle" in a reasonable amount of time we needed a systematic search method. Ad hoc procedures were simply too inefficient to be useful here. In fact, had we used ad hoc methods to find unretrieved relevant documents—that is, just "good guesses"—we would likely have found far fewer unretrieved, relevant documents and, more importantly, have estimated recall to be significantly higher.

Although the logical modification method was important in the STAIRS study as a systematic way to organize the search for unretrieved, relevant documents, it may, as we have said, be even more important as a basis for establishing comparable recall studies conducted on different retrieval systems. It would not, of course, give a "true" value for recall, but it would probably give a reasonable *maximum* value for recall that might be good enough to compare between different retrieval systems—something we have not had so far. In fact, one might argue that while a "true" recall value is theoretically possible, it is empirically elusive, leaving the method of recall estimation discussed here as one of the only candidates for reliable, comparable recall estimations on large systems. As Jordan commented on information retrieval evaluations, ". . . until the players in the game can agree on a set of rules for determining variables and evaluating test results, those of us who have to depend on their advice are not going to feel comfortable in doing so" (Jordan, 1989).

Logical Modification as a Search Procedure

The thoughtful reader may draw the correct inference that any good method for finding unretrieved relevant documents would also be effective as a search algorithm. This is particularly true for exhaustive (high recall) searches. In this sense, our logical modification/semantic expansion procedures could be a method for defining a series of queries that retrieve documents in a systematic way (in fact, the present version of this procedure had its origins in a query formulation procedure proposed in 1980 [Blair, 1980]). At the very least, since this method produces non-intersecting retrieved sets of documents, it insures that the searcher will at least be able to conduct his/her search without seeing the same documents repeatedly. At best, it defines a set of queries that systematically cover a widening search space beginning from the searcher's initial query. How far the searcher wishes to pursue the CCNF versions of the first query, is, of course, up to that individual.

This raises the natural question that if a searcher uses the logical modification method to conduct his/her search, how do you then find unretrieved relevant documents in any kind of systematic way? Clearly, if the search goes through *all* the candidate sets derivable from the initial query, or queries, our method provides no way to go beyond this (in effect, we are at the same point as we are with recall estimations today that do not use any systematic method). But, realistically, there would probably be few searches that would use *all* the candidate sets that could be generated, so those unexamined candidate sets could be examined for unretrieved, relevant documents.

The STAIRS Study: Its Allies and Its Rivals

The STAIRS study has attained a certain notoriety since its publication. Its results were so striking that it was difficult for anyone interested in document retrieval to remain neutral about its findings. Readers invariably formed strong opinions about the results of the study, opinions that ranged from enthusiastic support to outrage. The Information Retrieval community, as well as interested computer scientists were quickly polarized by our findings. Over the last decade the more common attitude towards the STAIRS study has been one of acceptance, no doubt because an increasing number of interested individuals has had first hand experience with large full-text retrieval systems like STAIRS—a rarity in the early–mid 1980s. The most prominent study to confirm our findings was that reported by Dabney, who found that the same poor retrieval results of full-text searching occurred with systems providing access to case law (it was thought, by detractors of the STAIRS study, that one of the reasons for the low recall levels that we found was the wide variety in the language that occurs in litigation support material. Language, it was argued, would be more "predictable" in case law and this would lead to higher recall values than what we found). Dabney's study touched off a spirited debate in *Law Library Journal* (see bibliographic section "Articles Which Discuss the STAIRS Evaluation in Some Detail").

Our (and, indirectly, Dabney's) results were further corroborated by internal studies done by Westlaw (personal communication). These findings were instrumental in convincing Westlaw to add descriptive information to supplement their full-text retrieval. (A theoretical discussion about why retrieval for case law has the same difficulties as for litigation support can be found in Blair, 1990 and 1995a).

Another empirical study, done recently, provided in-

direct corroboration for our results. Brooks, 1993 conducted an empirical test of an indexing strategy called "unlimited aliasing" (Furnas, Landauer, Gomez, & Dumais, 1987). Unlimited aliasing is essentially the same as full-text retrieval. But, as Brooks states: "This experiment found no evidence to support the strategy of unlimited aliasing . . . some index terms are simply better than others."

The most prominent criticism of the conclusions of the STAIRS study appeared in the more positive view of simple, full-text retrieval presented by Salton (1986). Our point-by-point response to Salton's arguments was published some time later in *Information Processing and Management* (Blair and Maron, 1990), after circulating for a number of years in the "invisible college" of information retrieval researchers.

While some readers may feel uncomfortable with the debate over the STAIRS study, such debate is the foundation of new knowledge. As Milton observed: "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, 1644).

Although the substance of this debate is of interest to anyone concerned with the validity of the STAIRS study, it will not be discussed here. Interested readers should consult the cited articles directly (see bibliographic section "The STAIRS Debate").

STAIRS and the Harvard Business Review Document Collection

The empirical findings of the STAIRS study are most frequently compared with the more optimistic results of full-text retrieval of articles published in the Harvard Business Review conducted by Tenopir (1985). As Jordan (1989) observed: ". . . these two studies, whose results are diametrically opposed, have become touchstones for protagonists of full text and surrogation." It is also the case that those who are in favor of full-text retrieval tend to cite the Tenopir study, while those who are more skeptical cite the STAIRS study. Yet these two studies' experimental design and goals are so divergent that any strict comparison of their results is simply not possible. (N.B., Tenopir herself does not draw these comparisons between her study and the STAIRS evaluation, this has been done by her readers.)

There are four major differences between the STAIRS study and the Tenopir evaluation:

1. Tenopir's study was conducted on a document collection of fewer than 1,000 documents, while the STAIRS study was conducted on a collection almost 40 times larger. The STAIRS study also includes a discussion of why the evaluations of small, full-text document retrieval performance are so much more optimistic than our results were.
2. Tenopir generated the evaluator's relevance judgments by using a panel of experts, not the individuals who originally formulated the queries. This method of relevance assessment was criticized, persuasively, by Swanson (1977) who argued that having a panel of experts judge relevance rather than the original searchers yields only "topical relevance," not utility. In the STAIRS study, we wanted to measure *utility* rather than *topicality* so all relevance judgments were made by the originators of the queries.
3. In the Tenopir study, only the results of the first set of queries submitted to the system were evaluated (each "set" consisting of four types of queries—full-text, abstract only, controlled vocabulary, and title only—on the same topic). Regardless of the results of these searches, the queries were not revised based on the success or failure of the original set of queries. In other words, the searches were not interactive. In the STAIRS study, we permitted the searchers to revise their original queries as many times as they liked, and to search until they believed that they had retrieved all the documents they wanted. No search consisted of a single query, and many queries went through 10 or more revisions, gathering more relevant documents during each iteration.
4. The Tenopir study calculated what is known as "relative recall." Relative recall was determined for a given query by comparing the number of relevant documents retrieved by that query to the total number of relevant documents retrieved by the union of the sets of documents retrieved by all four of the different kinds of queries for a given search. Relative recall studies compare the overlap of relevant documents retrieved by multiple queries, but do not spend any time searching for documents that were relevant but not retrieved by the original set of queries. In the STAIRS study, our primary goal was to find relevant documents that had been missed by the original search queries and all their iterations.

The values for relative recall are virtually certain to be higher than the recall values estimated by our methods. Since the Tenopir study did not spend any time searching for relevant documents that were not retrieved during the original searches, as was done in the STAIRS study, it cannot reject the hypothesis that it may have missed significant numbers of unretrieved relevant documents. Further, with no estimate of the "true" values of recall, one cannot say that the results of Tenopir's analysis are statistically significant. For example, Tenopir's study shows that, on average, more relevant documents are retrieved by the use of full-text searching than by any of the three other methods of searching (searching by the full-text of the abstracts alone, by controlled vocabulary, and by titles alone). These results are usually presented as percentages, and, as such, appear quite convincing—the mean recall values for the different search methods are:

Method	Mean recall
Full text	73.9%
Abstract	19.3%
Controlled vocabulary	28.0%
Bibliographic union	44.9%

These percentages indicate a striking advantage for full-text searching, and they are what are most often compared with the results of the STAIRS study which estimated the mean recall for full-text searches to be only 20%. But what is often left out, is that the double-digit recall advantage of full-text retreival is only a single-digit advantage when it comes down to the actual numbers of documents involved:

Method	Mean no. of relevant documents
Full-text	3.5
Abstract	1.0
Controlled vocabulary	1.2
Bibliographic union	2.0

Differences of only a document or two between search types, are not nearly as convincing as the same information represented in percentages (the STAIRS study, on the other hand, dealt with 100–200 relevant documents per search so it was less subject to the high variability characteristic of small sample sizes) but there is a further problem. If the *actual* total average number of relevant documents, retrieved and unretrieved, is significantly more than the number of relevant documents that exists in the union of the retrieved sets, then the differences between full-text searching and the three other ways of searching may not be statistically significant. For example, given the difference between mean full-text recall (3.5) and mean controlled vocabulary recall (1.2), it looks like full-text retrieval retrieves almost three times the number of relevant documents that controlled vocabulary searching does. But, if the total *average* number of documents relevant to the searches is, say, 50 or 60, then a 2.3 (i.e., 3.5–1.2) document difference is not statistically significant. Without some estimate of the actual recall values, Tenopir's claim for the superiority of full-text searching over the other search methods is not supported. (Curiously, the published version of Tenopir's study lacks even an estimation of the standard deviation for the recall values she discovered. In the STAIRS study our standard deviation for recall was 15.9%, so even in a best-case scenario, the mean STAIRS recall value would be 35.9% (15.9 + 20.0), a result that is still consistent with the findings we reported and the conclusions we drew.)

It is also easy to show that there is a very high likelihood that Tenopir missed a significant number of relevant documents in her searches—that is, that the total number of relevant documents retrieved by the union of all the retrieved sets is substantially fewer than the total number of actual relevant documents that exists in the document collection. All the searches that Tenopir conducted were specified ahead of time, to remove any bias that might develop as she got more and more familiar with the document collection—a problem with searching on a small document collection. The difficulty with this approach is that it puts an artificial constraint on the searching process and reduces the likelihood of finding a significant proportion of relevant documents with the original queries. As Swanson (1977) has argued, information retrieval is a trial and error process. A typical search usually consists of multiple queries, with the construction of subsequent queries informed by the success or failure of previous queries. The likelihood that a searcher would get a substantial portion of the relevant documents when they are not permitted to revise their original query is, we think, remote. This means that it is likely that the Tenopir study missed a significant proportion of the total relevant documents in the collection, and, further, that the differences she found between different search methods are not statistically significant. Tenopir's single-query approach may be the reason why there were so many searches in her study which returned *no* relevant documents. In the STAIRS study we also had *initial* queries that returned no relevant documents, but we never had a complete search that returned no relevant documents. Further, in the STAIRS study, our searchers never finished a search with just a single query.

We have observed (above) that the biggest problem with the comparability of recall studies is that they have a number of significant uncertainties. It is possible to reduce three of these uncertainties (a definition of relevance for the searchers, a definition of relevance for the evaluators, and a stopping rule for the searchers) in the manner we described. But until the advent of the STAIRS study there was no way to reduce the other two uncertainties (where should the evaluators look for unretrieved, relevant documents, and when should they stop looking for them) in any kind of systematic way. Logical modification goes a long way towards reducing these two indeterminacies. The only prerequisite is that the queries submitted to the system must be convertible to complete, conjunctive normal form. Although this may seem to be a difficult prerequisite, any query that can be represented in propositional or Boolean logic can be converted, without loss, to complete conjunctive normal form. Even queries written in natural language are often convertible into propositional logic, as any logic text will show.

Why Retrieval Using STAIRS Should Have Been Better

Many readers of the STAIRS study wondered why STAIRS did so poorly, especially considering that the recall values calculated were *maximum* values—the "true" values were undoubtedly even lower than the 20% mean that we observed. But what is even more striking about the STAIRS study is that the environment for retrieval was particularly auspicious—much more favorable than could be expected on a "typical" document retrieval system of that size. The reason for this is that the

documents STAIRS provided access to were personally selected (from a larger set of documents) by the two lawyers and two paralegals who participated in the study. This selection process spanned a 1-year period and produced a set of documents that were all germane to the various issues in the lawsuit. This meant that the searchers who participated in the STAIRS study had seen and selected each of the documents in the collection. In effect, STAIRS was being used to manage a *personal* document collection. In this context it is even more striking that the recall levels were so far below what the lawyers had expected. Further, if recall levels for a "personal" information retrieval system such as the STAIRS system were so low, how much less auspicious would retrieval be on a system where the searchers were not familiar with the documents in the collection?

This raises several important issues: In the first place, how could the lawyers' recollection of the documents relevant to their queries (and which they had already seen at least once) be so poor? There were times during the course of a particular search when the lawyers said they recalled more relevant documents than they found, but they would just continue searching and always finished their searches with the belief that they had gotten at least 75% of them, as they had stipulated. One of the reasons for this inability to retrieve many of the previously seen, relevant documents was that the lawyers could not recall the exact words which occurred uniquely in them. (By "uniquely" we mean those words or phrases that occurred in the relevant documents, but did *not* occur in non-relevant ones—a fundamental requirement for effective retrieval on systems with many documents.) What they remembered was the "gist" of the documents they had seen before, but neither the exact number of those relevant documents, nor the precise wording that uniquely occurred in them. Psychologists have demonstrated convincingly that a subject's recollection of things past is typically not literal, no matter how important those recollections might be (Barclay, Bransford, & Franks, 1972; Brewer, 1975; Fillenbaum, 1966; Just and Carpenter, 1976; Levelt and Kempen, 1975; Sachs, 1967; Wanner, 1974; inter alios). Yet, we might venture, the implicit assumption of simple full-text retrieval systems like STAIRS is that we all have this literal recall ability (if we are searching for documents that we have seen already). In fact, we do not. Because of this, we not only will not have very good anticipation of the exact words and phrases that occur uniquely in textual passages that we might want, we do not even have good recall of the exact words and phrases of those documents *that we have already seen and want again.* (To get a feel for this, the reader is invited to recollect the number of times (both when and where) he/she used a familiar word in the last week. Most readers, we imagine, would find these quite difficult to do.) To a psychologist, what we found in the STAIRS study should not be too surprising, and to the information retrieval researcher the implication of the STAIRS study is that simple full-text retrieval presupposes a cognitive ability—literal recall or anticipation of words that uniquely occur in relevant documents—that most people do not have, even unusually capable subjects like the lawyers in our test. These lawyers were successful partners in a major corporate law firm so it would be unlikely that their failure to predict the words and phrases in the documents they wanted was a result of some inferior intellectual ability. In fact, the practice of law often places exceptionally high demands on lawyers' ability to recall important literal information—such as verbatim testimony. We might expect that the average searcher would not have even this capacity.

STAIRS' Enhancements to Simple Full-Text Retrieval

Some results of the STAIRS study that have not been reported before are the tests of STAIRS' ranking algorithms and its automatic thesaurus. STAIRS had a set of five document ranking algorithms based on different word frequency calculations (see Table 1). These were not complex algorithms, but one might expect that they would at least provide a marginal improvement over the unordered retrieved sets of documents. This turned out not to be the case. There was no statistically significant correlation between the rank ordering of the retrieved sets by any of the five algorithms and the ranking that the users' relevance judgments placed on those retrieved sets. While these are only five out of myriad such statistical ranking procedures, their inability to predict the relevance ranking of the retrieved documents means that there is no simple automatic solution to the problem of improving full-text retrieval systems. It also means that we cannot reject the hypothesis that an experienced searcher, using simple full-text searching techniques and looking through relatively familiar material can have search results that are just as good as, if not better than, searches augmented by simple word frequency calculations.

STAIRS also had an on-line thesaurus, called the TLS (*T*hesaurus *L*inguistic *S*ystem). This was a manually constructed thesaurus that could be used to provide synonyms for search terms. It was constructed over 18

TABLE 1. STAIRS document ranking algorithms.[†]

1. $D_V = (F_{TD} * F_T)/T_T$
2. $D_V = F_{TD}$
3. $D_V = (F_{TD} * F_{TD})/T_T$
4. $D_V = (F_{TD} * F_T)/(T_T + T_T)$
5. $D_V = (F_{TD} * T_T)/(F_T - F_{TD})$

[†] D_V = Document value; F_{TD} = The number of occurrences of term T in document D; F_T = The number of occurrences of term T in the retrieved set; T_T = The number of documents in the retrieved set in which term T occurs.

months (at a cost of about $150,000) by an engineer who worked full-time on the project. Yet, over the course of the STAIRS evaluation, not a single relevant document (retrieved or unretrieved) was found with the TLS that had not been found with the simple full-text searching techniques of STAIRS. The reason for this was abundantly clear. The thesaurus linked semantically related engineering terms, but these semantic relations were not really that useful for searching in the lawsuit. What the lawyers needed were synonymous ways in which the *legal* issues were discussed, and this was very difficult to predict ahead of time. (The original Blair and Maron (1985) article gives some examples of the ad hoc term correlations that occurred in the STAIRS study.)

Was the STAIRS System Itself an Anomaly?

One question about the results of the STAIRS study concerned the characteristics of the system itself and the searching ability of the lawyers and paralegals. Specifically, it could have been argued that the document retrieval system we studied may have been anomalous in a way that predisposed it towards poor recall levels—that, for example, the searchers were inexperienced, or the document collection unusually difficult to search through. Yet it was IBM itself which dispelled this objection. IBM frequently brought potential STAIRS customers to see the system we were studying. They also used the lawyers and paralegals who participated in the evaluation to demonstrate searching on the system to IBM's potential STAIRS customers (highly sensitive information was excluded from these searches). So, in spite of the difficulties that we were to find with the STAIRS system, IBM considered it an exemplary system whose searchers—the ones who participated in our study—were considered better demonstrators of STAIRS than IBM's own representatives.

A further indication that the searchers were operating at the best of their ability throughout the test was the lack of evidence for any learning curve on their part (see Blair, 1990). Searches done during the first half of the test had the same mean level of success that searches conducted during the second half did.

It appears that, in spite of the low level of retrieval that we observed with STAIRS, the retrieval environment that STAIRS operated in was unusually propitious, and that simple full-text retrieval in other environments would likely perform at significantly lower levels. Instead of the 20% recall, we observed being a "worst" case, or even an "average" case, it appears, on full reflection, that what we found was clearly a "best" case level of document retrieval for STAIRS.

Our Strong Denunciation of Simple Full-Text Retrieval

Many readers, while sympathetic with our results, were concerned by our strong denunciation of simple full-text retrieval systems, feeling that such systems were better than nothing. But our low estimation of simple full-text retrieval had a different context than it does today. In the late 1970s, when the study was done, word processing systems were practically unknown in business, so that getting a machine readable document required that the paper copy of the document be typed into STAIRS and verified for accuracy. For STAIRS the cost of this process was $26.00 per document. Since there were about 40,000 documents that had been entered into the system by the time of the study, the cost of input alone was $26 × 40,000, or $1,040,000. (It was anticipated that the final number of documents that STAIRS would manage if the lawsuit made it to court, was about 1.5 million documents, making the data entry cost alone almost $40 million.) To our minds, this, clearly, was too great a price to pay for the "advantage" of full-text retrieval. With that kind of "up front" cost for simple full-text retrieval, we felt that it was certainly the case that, as we quoted Dr. Johnson, ". . . one is surprised to see it done at all." At the time of the STAIRS study, the best large-scale document retrieval systems contained primarily hard-copy versions of the documents or texts. Some, such as the Library of Congress contained upwards of 50 million items or texts. But the up front costs and complexities of indexing made construction of these kinds of systems discouraging for managers not familiar with traditional methods of information retrieval. It appeared that STAIRS might be a "magic bullet," enabling you to bypass this complex indexing process. To businesses, it seemed like a good tradeoff. But the manually indexed information retrieval systems that we observed in industry at that time, were far less costly to set up than STAIRS. For an experienced indexer, indexing technical documents was a much faster process than typing in an entire 10 page document. So the input of documents for manually indexed systems was actually faster and cheaper than the full-text document representations that STAIRS required. Although we did not make a comparison of the retrieval effectiveness of these manually indexed systems with STAIRS, it would be doubtful that their recall values would have been much worse than the *maximum* 20% recall value we observed with STAIRS. That, coupled with faster, cheaper setup times made STAIRS an expensive alternative that could not guarantee improved results.

Today, the "better than nothing" argument for simple full-text retrieval is more convincing than it was 10 or 15 years ago. Most documents now *begin* as machine readable documents, so there is no up front cost of typing the documents into an information system. Within this context, a simple full-text retrieval system is arguably better than nothing, and cheaper than a system using manually or automatically assigned index terms. But our other caveats remain, namely, that for an information retrieval system to be used for a "mission critical" or, high recall, application, the capabilities of sim-

ple full-text retrieval alone are just not up to the task. In such systems, the simple full-text retrieval must be augmented with a carefully thought out logical or intellectual structure, usually based on the activity that those documents serve (Blair, 1990, 1995b). In the STAIRS study the obvious logical structure that could have been used was the written complaint on which the lawsuit was based. There were 13 specific issues described in the complaint. Each document was germane to only one issue. If each of the STAIRS documents had been assigned an indexing term relating it to the one specific issue it was concerned with, the 40,000 document data base would have been partitioned into 13 smaller non-intersecting document collections. Each of these 13 document collections would naturally be much smaller than the 40,000 document collection that we had to use. Since one of our main arguments in the STAIRS study was that recall values fall off significantly as the document collection grows larger, any strategy that partitions a large document collection into a number of smaller ones would likely improve the recall capability of the system (Blair, 1995a).

There are some document collections and query types that are particularly suited to simple full-text retrieval, and might have higher recall rates than we observed. These systems contain a lot of precise contextual references and proper names that can be easily identified. An example of such a document collection would be a collection of newspaper articles (such as the *New York Times* or *Wall Street Journal*). News articles are written with particular care to report proper names, dates, etc. very accurately. For example, articles about Henry Kissinger, or IBM will most certainly have these proper names in them, and articles that do not contain them would not be likely to be relevant to a search for them. Journalists are urged to write their articles carefully including and verifying such facts. Many retrieval requests submitted to newspaper document collections are looking for articles with these kind of proper nouns or dates in them. In such a situation, recall values would likely be higher than what we observed. Of course there are some situations where the searcher might want to find news articles that do not have obvious specific terms in them—for example, a request for recent articles that discuss religious cults, or articles that discuss Southeast Asian political issues. In these situations a simple full-text retrieval system would probably not be able to perform higher than the rate we found in the STAIRS study, ceteris paribus.

Some Final Thoughts

As much as we would like commercial applications of information retrieval to be successful, it appears that we, as information retrieval researchers, have a long ways to go if we are ever to build successful "mission-critical" information retrieval systems. Some might say that "mission-critical" information retrieval systems are too rare to worry about, and that the low recall requirements of average systems are probably well within our reach. But this is not a valid objection. In the first place, the need for mission-critical retrieval is growing rapidly both in numbers and in the percentage of commercial systems. This rise is due to the increasing use of document retrieval to support critical decision-making processes within organizations (Blair, 1995b). These critical systems *will* be built, whether information retrieval researchers choose to participate or not. Further, even if non-critical systems are our primary focus, it is still important for us to understand the fundamental dynamics of information retrieval and to build these modest systems with the best and most effective retrieval mechanisms possible—in the same way that while the needs of the average motorist do not demand a car that can win the Daytona 500, the auto manufacturers' building of such high performance cars informs the design and manufacture of their "average" cars.

Document retrieval has long been the poor stepchild of the computer revolution, with the lion's share of investment and research funding going to the development of data retrieval systems. But this is changing, especially since it has become clear that advances in document retrieval cannot be leveraged off advances in data retrieval (Blair, 1995a). Data retrieval and document retrieval are two different activities, requiring two different models. In fact, it is our contention that the basic model for information management, when it concerns both data and documents, is *not* the data model, but the document retrieval model. Basically, the data retrieval model is a restricted subset of the document retrieval model—data retrieval being like document retrieval without the inherent uncertainty of description and query formulation.

Some businessmen are genuinely enthusiastic about the "new" field of document management. In a front page *Wall Street Journal* article, William Lowe, the force behind IBM's personal computer and then vice president at Xerox, hailed computerized management of documents as ". . . the big news of the 1990s in much the same way personal computers were the big news of the 1980s" (Hooper, 1990). Commercial interests are finally coming to realize that the data model does not fit the document retrieval model well, and businesses are finding that the majority of the information that they keep is embedded in documents. Documents are where data becomes knowledge, and, in some sense, we might venture that the intelligence of an organization—its organizational memory—exists in its documents rather than its data bases. If organizations cannot provide reasonably good access to this textual information, they run the risk of management by amnesia—of not being able to "remember" its past triumphs (and to build on them) nor its past failures (and to avoid doing them again).

There is an urgency, now, in information retrieval

design, primarily because the task of information retrieval is becoming increasingly difficult. Previously, the cost of storing textual information was a significant component of the cost of information management: Most textual information was in some kind of hard copy form, and hard copy storage is costly. In downtown San Francisco in the late 1970s (when the STAIRS study was done), the cost of office space needed to keep a single filing cabinet was around $200 per month. While this appeared to be a "cost" of information management, in reality it was a benefit. Since there was limited space for hard copy storage, when the space ran out, you had to "weed" out the less important information that you had, to make room for the important things you wanted to keep. True, the cost of computer storage was well below the cost of hard copy storage even then, but most textual information started as hard copy, so the cost of converting hard copy to machine-readable form was significant enough to discourage the conversion of all but the most important information. But now most texts, images, graphics, etc., *begin* in machine-readable format, and with the cost of computer storage continuing to fall dramatically, it is now practical for even the largest company to keep everything it has ever written—from documents detailing corporate strategy, to memoes about the office bowling league. The low cost of information storage has created an "information landfill" in most companies (Blair, 1995b). By keeping every memo, trivial or important, that was ever written in an organization we are subject to two problems: First, our document collections are going to be larger—perhaps, dramatically larger—than they ever were. Since our claim in the STAIRS study was that document retrieval performance degrades as document collections get larger, document retrieval is becoming harder now than it used to be, merely because the size of a typical document collection is so much larger than it used to be. Although there is an increasing number of commercial document retrieval tools available, it is a plausible hypothesis that document retrieval effectiveness may be getting worse faster than we are improving it, and without comparable recall studies, we may not even know how great this deficit is.

The second problem with an organization keeping all its documents is that the searcher must wade through the non-relevant, perhaps even trivial, information in order to get access to the important documents. Trivial, useless information is just "noise" in a document retrieval system, and given the low levels of effectiveness that most document retrieval systems probably run at, we cannot afford to degrade performance even further.

This brings us to two of the central questions of this issue of *JASIS*—how well do we evaluate document retrieval performance, and how can we improve these techniques? Of fundamental importance is the fact that information retrieval research has no standard test of retrieval effectiveness for intellectual access that can be used to compare different systems' performance. If we do not make substantial progress in finding a standard measurement of document retrieval performance, then we will not be able to distinguish between less effective and more effective retrieval techniques. We may then be faced with a future in which our research will not be convincing enough to justify investment in its application. Lacking commercial support, information retrieval research may then be reduced to a quiet intellectual backwater of information management, known only for its interesting puzzles that demonstrate a high degree of rigor but a low degree of relevance to operational systems. The first quotation by Jordan (end of "Semantic Expansion" section, above) should be a warning to us all. There are many fascinating puzzles and issues in information retrieval, but there must be at least *some* practical component to our work. Commercial information retrieval vendors began attending our information retrieval conferences in significant numbers in the late 1980s. Like Jordan, they are looking for answers. It is at least part of our responsibility to attempt to provide some of those answers—TREC being a good example of how to bridge this link between theory and application. This is not to say that all our work should satisfy practical, commercial demands—that would be to err at the other extreme. Perhaps our fundamental juxtaposition is between theory and practice. Both can inform each other, but, as Kuhn noted, scientific disciplines speak their own dialects, and we in information retrieval have our own way of talking about the things that interest us (Kuhn, 1970). If the representatives of commercial document management interests come to our conferences and it sounds to them like we researchers are speaking in "tongues," then there will be no common ground on which to build better systems. Meanwhile, the long-term trend is for document collections to get larger and larger, making effective, high recall retrieval increasingly difficult. If it *is* the case that performance is getting worse faster than we can improve it, then this is a fate that none of us either deserves or wants.

Acknowledgments

I acknowledge the enormous help that M. E. Maron provided through his extensive comments on earlier versions of this article.

References

Barclay, J. R., Bransford, J. D., & Franks, J. J. (1972). Sentence memory: A constructive versus interpretive approach. *Cognitive Psychology, 3*. 193–209.

Bauman, N. (1994). The illusions and realities of full-text searching. *Law Office Computing, 4*(3), 44–52.

Blair, D. C. (1980). Searching biases in large, interactive document

retrieval systems. *Journal of the American Society for Information Science, 31,* 271–277.
Blair, D. C. (1982). The Nature of Scientific Theory, *Human Systems Management, 3,* 279–288.
Blair, D. C. (1984a). The data-document distinction in information retrieval. *Communications of the ACM, 27*(4), 369–374.
Blair, D. C. (1984b). The management of information: Basis distinctions. *Sloan Management Review, 26*(1), 13–23.
Blair, D. C. (1986). Indeterminacy in the subject access to documents. *Information Processing and Management, 22*(2), 229–241.
Blair, D. C. (1988). An extended relational document retrieval model. *Information Processing and Management, 24*(3), 349–371.
Blair, D. C. (1990). *Language and representation in information retrieval.* Amsterdam: Elsevier Science.
Blair, D. C. (1995a). *The challenge of document retrieval: Major issues and a framework based on search exhaustivity and data base size.* Unpublished working paper.
Blair, D. C. (1995b). *The revolution in document management: Corporate memory or information landfill?* Unpublished working paper.
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), 289–299.
Blair, D. C., & Maron, M. E. (1990). Full-text information retrieval: Further analysis and clarification. *Information Processing and Management, 26,* 437–447.
Brewer, W. F. (1975). Memory for ideas: Synonym substitution. *Memory and Cognition, 3,* 458–464.
Brooks, T. A. (1993). All the right descriptors: A test of unlimited aliasing. *Journal of the American Society for Information Science, 44,* 137–147.
Cooper, W. S. (1973). On selecting a measure of retrieval effectiveness. *Journal of the American Society for Information Science, 24,* 87–100.
Copi, I. (1965). *Symbolic logic* (2nd ed.). New York: Macmillan and Company.
Delphi Consulting Group, Inc. (1992) *Text retrieval systems: A market and technology assessment.* (Available from 266 Beacon St., Boston, MA 02116-1224).
Fillenbaum, S. (1966). Memory for gist: Some relevant variables. *Language and Speech, 9,* 217–227.
Fleischer, R. (1990). Total document control: A text-retrieval perspective. In P. Gillman (Ed.), *Text retrieval: Information first. Proceedings of the Institute of Information Scientists 1990 Text Retrieval Conference, London, October 1990.* London: Taylor Graham.
Furnas, G. W., Landauer, T. K., Gomez, L. M., & Dumais, S. T. (1987). The vocabulary problem in human-system communication. *Communications of the ACM, 30*(11), 964–971.
Gey, F. & Dabney, D. (1990). [Letters to the editor]. *Journal of the American Society for Information Science, 40,* 613.
Hooper, L. (1990, September 20). High-tech gamble: Xerox tries to shed its has-been image with big new machine. *Wall Street Journal,* p. 1.
Jordan, J. (1989). [Letters to the editor.] *Journal of the American Society for Information Science, 40,* 362–363.
Just, M. A., & Carpenter, P. A. (1976). The relation between comprehending and remembering some complex sentences. *Memory and Cognition, 4,* 318–322.
Kuhn, T. (1970). *The structure of scientific revolutions* (2nd ed.). Chicago: University of Chicago Press.
Levelt, W. J. M., & Kempen, G. (1975). Semantic and syntactic aspects of remembering sentences: A review of some recent continental research. In A. Kennedy & A. Wilkes (eds.), *Studies in long term memory* (pp. 201–218). London.
Milton, J. (1644) *Areopagitica.*
Paijmans, H. (1993). Comparing the document representations of two IR systems: CLARIT and TOPIC. *Journal of the American Society for Information Science, 44,* 383–392.
Popper, K. (1959). *The logic of scientific discovery.* New York: Basic Books.
Popper, K. (1968). *Conjectures and refutations.* New York: Harper and Row.
Sachs, J. S. (1967). Recognition memory for syntactic and semantic aspects of connected discourse. *Perception and Psychophysics, 2,* 437–442.
Salton, G. (1986). Another look at automatic text-retrieval systems. *Communications of the ACM, 29*(7), 648–656.
Salton, G. (1989). *Automatic text processing.* Reading, MA: Addison-Wesley.
Sembok, T. M. T., & van Rijsbergen, C. J. (1990). Silol: A simple logical-linguistic document retrieval system. *Information Processing and Management, 26,* 111–134.
Swanson, D. R. (1960). Searching natural language text by computer. *Science, 132,* 1960–1104.
Swanson, D. R. (1977). Information retrieval as a trial and error process. *Library Quarterly, 47*(2), 128–148.
Tague, J. (1981). The pragmatics of information retrieval experimentation. In K. Sparck Jones (Ed.), *Information Retrieval Experiment* (chap. 5). London: Butterworths.
Tenopir, C. (1985). Full text database retrieval performance. *On-line Review, 9,* 149–164.
van Rijsbergen, K. (1979). *Information retrieval* (2nd ed.). London: Butterworths.
Wanner, E. (1974). *On remembering, forgetting, and understanding sentences.* The Hague: Mouton.

Bibliography of Articles Relating to the STAIRS Study

Articles Detailing the STAIRS Study

Blair, D. C. (1986). Full-text retrieval: Evaluation and implications. *International Journal of Classification, 13*(1), 18–23.
Blair, D. C. (1990). *Language and representation in information retrieval.* Amsterdam: Elsevier.
Blair, D. C., & Maron, M. E. (1985). An evaluation of retrieval effectiveness for a full-text document retrieval system. *Communications of the ACM, 20,* 289–299.

The STAIRS Debate

Blair, D. C., & M. E. Maron (1990). Full-text information retrieval: Further analysis and clarification. *Information Processing and Management, 26,* 437–447.
Blair, D. C., & Maron, M. E. (1985). Technical Correspondence. *Communications of the ACM, 28*(11), 1238–1242.
Salton, G. (1986). Another look at automatic text-retrieval systems. *Communications of the ACM, 20,* 648–656.

Articles which Discuss the STAIRS Evaluation in Some Detail

Bauman, N. (1994). The illusions and realities of full-text searching. *Law Office Computing, 4*(3), 44–52.
Berring, R. C. (1986). Full-text databases and legal research: Backing into the future. *High Technology Law Journal, 1.*
Bing, J. (1987). Performance of legal text retrieval systems: The curse of Boole. *Law Library Journal, 79,* 187–202.
Burson, S. F. (1987). A reconstruction of Thamus—comments on the evaluation of legal information retrieval systems. *Law Library Journal, 79*(1), 133–143.

Dabney, D. P. (1986). The curse of Thamus: An analysis of full-text document retrieval. *Law Library Journal, 78*(1), 5–40.

Dabney, D. P. (1986). West Publishing Company and Mead Data Central on the curse of Thamus—a reply. *Law Library Journal, 78*(2), 349–350.

Doyle, J. (1990). Aiming at the databases: Retrieval effectiveness of legal databases. *Trends in Law Library Management and Technology, 3.*

Gey, F., & Dabney, D. (1990). [Letters to the editor]. *Journal of the American Society for Information Science, 41,* 613–614.

Jordan, J. (1989). [Letters to the editor]. *Journal of the American Society for Information Science, 40,* 613.

McDermott, J. (1986). Another analysis of full-text legal document-retrieval. *Law Library Journal, 78*(2), 331–340.

Runde, C. E., & Lindberg, W. H. (1986). The curse of Thamus—A response. *Law Library Journal, 78*(2), 345–347.

Salton, G. (1992). The state of retrieval-system evaluation. *Information Processing and Management, 28*(4), 441–449.

Sutton, S. A. (1994). The role of attorney mental models of law in case relevance determinations—an exploratory analysis. *Journal of the American Society for Information Science, 45,* 186–200.

Swanson, D. R. (1988). Historical note—information retrieval and the future of an illusion. *Journal of the American Society for Information Science, 39,* 92–98.

Voges, M. A. (1988). Information systems and the law. *Annual Review of Information Science and Technology, 23,* 193–216.

Wren, C. & Wren, J. (1994). *Using computers in legal research: A guide to LEXIS and WESTLAW.* Madison, Wisconsin: Adams and Ambrose Publishing.

Zoellick, W. (1986). Selecting an approach to document retrieval (chapter 5). In S. Ropiequet (Ed.), *CDROM the new papyrus,* (Vol. 2, pp. 63–83). Seattle: Microsoft Press.