Edgar H. Sibley
Panel Editor

An evaluation of a large, operational full-text document-retrieval system (containing roughly 350,000 pages of text) shows the system to be retrieving less than 20 percent of the documents relevant to a particular search. The findings are discussed in terms of the theory and practice of full-text document retrieval.

AN EVALUATION OF RETRIEVAL EFFECTIVENESS FOR A FULL-TEXT DOCUMENT-RETRIEVAL SYSTEM

DAVID C. BLAIR and M. E. MARON

Document retrieval is the problem of finding stored documents that contain useful information. There exist a set of documents on a range of topics, written by different authors, at different times, and at varying levels of depth, detail, clarity, and precision, and a set of individuals who, at different times and for different reasons, search for recorded information that may be contained in some of the documents in this set. In each instance in which an individual seeks information, he or she will find some documents of the set useful and other documents not useful; the documents found useful are, we say, *relevant*; the others, not relevant.

How should a collection of documents be organized so that a person can find all and only the relevant items? One answer is automatic full-text retrieval, which on its surface is disarmingly simple: Store the full text of all documents in the collection on a computer so that every character of every word in every sentence of every document can be located by the machine. Then, when a person wants information from that stored collection, the computer is instructed to search for all documents containing certain specified words and word combinations, which the user has specified.

Two elements make the idea of automatic full-text retrieval even more attractive. On the one hand, digital technology continues to provide computers that are larger, faster, cheaper, more reliable, and easier to use; and, on the other hand, full-text retrieval avoids the need for human indexers whose employment is increasingly costly and whose work often appears inconsistent and less than fully effective.

A pioneering test to evaluate the feasibility of full-text search and retrieval was conducted by Don Swanson and reported in *Science* in 1960 [6]. Swanson concluded that text searching by computer was significantly better than conventional retrieval using human subject indexing. Ten years later, in 1970, Salton, also in *Science*, reported optimistically on a series of experiments on automatic full-text searching [3].

This paper describes a large-scale, full-text search and retrieval experiment aimed at evaluating the effectiveness of full-text retrieval. For the purposes of our study, we examined IBM's full-text retrieval system, STAIRS. STAIRS, an acronym for "**ST**orage **A**nd **I**nformation **R**etrieval **S**ystem," is a very fast, large-capacity, full-text document-retrieval system. Our empirical study of STAIRS in a litigation support situation showed its retrieval effectiveness to be surprisingly poor. We offer theoretical reasons to explain why this poor performance should not be surprising and also why our experimental results are not inconsistent with the earlier more favorable results cited above. The retrieval problems we describe would be problems with any large-scale, full-text retrieval system, and in this sense our study should not be seen as a critique of STAIRS alone, but rather a critique of the principles on which it and other full-text document-retrieval systems are based.

THE ALLURE OF FULL-TEXT DOCUMENT RETRIEVAL

Retrieving document texts by subject content occupies a special place in the province of information retrieval because, unlike data retrieval, the richness and flexibility of natural language have a significant impact on the conduct of a search. The indexer chooses subject terms that will describe the informational content of the documents included in the database, and the user describes his or her information need in terms of the subject descriptors actually assigned to the documents (Figure 1). However, there are no clear and precise rules to govern the indexers' choice of appropriate subject terms, so that even trained indexers may be inconsistent in their application of subject terms. Experimental studies have demonstrated that different indexers will generally index the same document differently [9], and even the same individual will not always select the identical index terms if asked at a later time to index a document he or she has already indexed. The problems associated with manual assignment of subject descriptors make computerized, full-text document retrieval extremely appealing. By entering the entire, or the most significant part of, a document text onto the database, one is freed, it is argued, from the inherent evils of manually creating document records reflecting the subject content of a particular document; among these, the construction of an indexing vocabulary, the training of indexers, and the time consumed in scanning/reading documents and assigning context and subject terms. The economies of full-text search are appealing, but for it to be worthwhile, it must also provide satisfactory levels of retrieval effectiveness.

MEASURING RETRIEVAL EFFECTIVENESS

Two of the most widely used measures of document-retrieval effectiveness are Recall and Precision. Recall measures how well a system retrieves *all* the relevant documents; and Precision, how well the system retrieves *only* the relevant documents. For the purposes of this study, we define a document as relevant if it is judged useful by the user who initiated the search. If not, then it is nonrelevant (see [4]). More precisely, Recall is the proportion of relevant documents that the system retrieves, the ratio of x/n_2 (Figure 2). Notice that one can interpret Recall as the probability that a relevant document will be retrieved. Precision, on the other hand, measures how well a system retrieves *only* the relevant documents; it is defined as the ratio x/n_1 and can be interpreted as the probability that a retrieved document will be relevant.

THE TEST ENVIRONMENT

The database examined in this study consisted of just under 40,000 documents, representing roughly 350,000 pages of hard-copy text, which were to be used in the

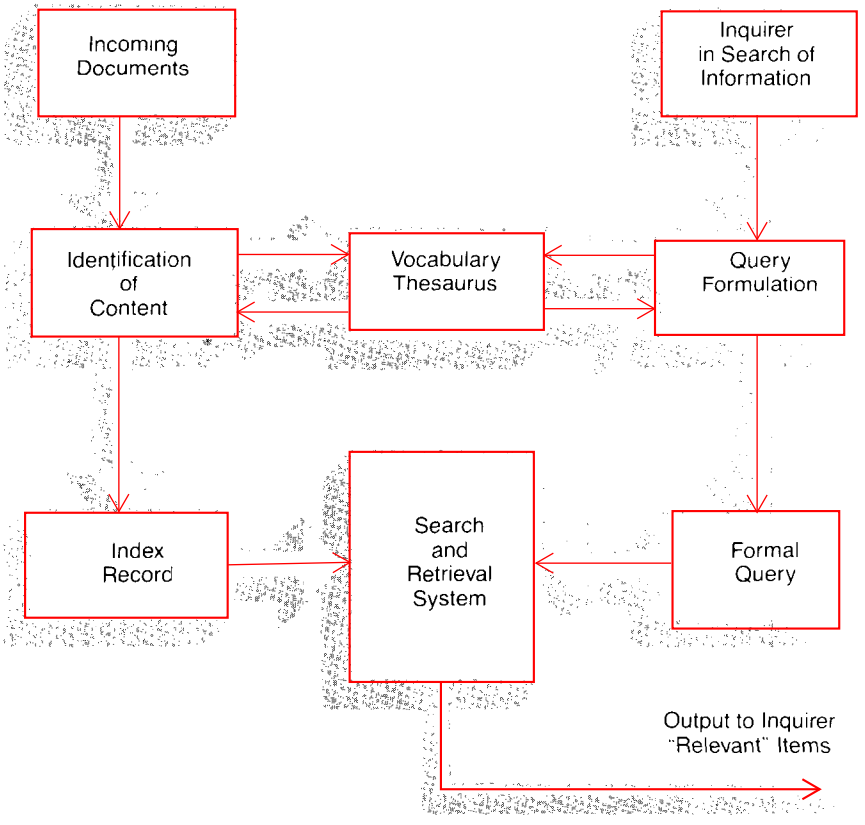

FIGURE 1. The Dynamics of Information Retrieval

$$\text{Recall} = \frac{\text{Number of Relevant and Retrieved}}{\text{Total Number Relevant}} = \frac{x}{n_2}$$

$$\text{Precision} = \frac{\text{Number of Relevant and Retrieved}}{\text{Total Number Retrieved}} = \frac{x}{n_1}$$

FIGURE 2. Definitions of Precision and Recall

defense of a large corporate law suit. Access to the documents was provided by IBM's STAIRS/TLS software (**ST**orage **A**nd **I**nformation **R**etrieval **S**ystem/**T**hesaurus **L**inguistic **S**ystem). STAIRS software represents state-of-the-art software in full-text retrieval. It provides facilities for retrieving text where specified words appear either singly or in complex Boolean combinations. A user can specify the retrieval of text in which words appear together anywhere in the document, within the same paragraph, within the same sentence, or adjacent to each other (as in "New"adjacent "York"). Retrieval can also be performed on fields such as author, date, and document number. STAIRS provides ranking functions that permit the user to order retrieved sets of 200 documents or less in either ascending or descending numerical (e.g., by date) or alphabetic (e.g., by author) order. In addition, retrieved sets of less than 200 documents can also be ordered by the frequency with which specified search terms occur in the retrieved documents. The Thesaurus Linguistic System (TLS) provides the facilities to manually create an interactive thesaurus that can be called up by the user to semantically broaden (or narrow) his or her searches; it allows the designer to specify semantic relationships between search terms such as "narrower than," "broader than," "related to," "synonomous with," as well as automatic phrase decomposition. STAIRS/TLS thus represents a comprehensive full-text document-retrieval system.

THE EXPERIMENTAL PROTOCOL

To test how well STAIRS could be used to retrieve *all* and *only* the documents relevant to a given request for information, we wanted in essence to determine the values of Recall (percentage of relevant documents retrieved) and Precision (percentage of retrieved documents that are relevant). Although Precision is an important measure of retrieval effectiveness, it is meaningless unless compared to the level of Recall desired by the user. In this case, the lawyers who were to use the system for litigation support stipulated that they must be able to retrieve at least 75 percent of all the documents relevant to a given request for information, and that they regarded this entire 75 percent as essential to the defense of the case. (The lawyers divided the relevant retrieved documents into three groups: "vital," "satisfactory," and "marginally relevant." All other retrieved documents were considered "irrelevant.")

CONDUCT OF THE TEST

For the test, we attempted to have the retrieval system used in the same way it would have been during actual litigation. Two lawyers, the principal defense attorneys in the suit, participated in the experiment. They generated a total of 51 different information requests, which were translated into formal queries by either of two paralegals, both of whom were familiar with the case and experienced with the STAIRS system. The paralegals searched on the database until they found a set of documents they believed would satisfy one of the initial requests. The original hard copies of these documents were retrieved from files, and xerox copies were sent to the lawyer who originated the request. The lawyer then evaluated the documents, ranking them according to whether they were "vital," "satisfactory," "marginally relevant," or "irrelevant" to the original request. The lawyer then made an overall judgment concerning the set of documents received, stating whether he or she wanted further refinement of the query and further searching. The reasons for any subsequent query revisions were made in writing and were fully recorded. The information-request and query-formulation procedures were considered complete only when the lawyer stated in writing that he or she was satisfied with the search results for that particular query (i.e., in his or her judgment, more than 75 percent of the "vital," "satisfactory," and "marginally relevant" documents had been retrieved). It was only at this point that the task of measuring Precision and Recall was begun. (A diagram of the information-request procedure is given in Figure 3.) The lawyers and paralegals were permitted as much interaction as they thought necessary to ensure highly effective retrieval. The paralegals were able to seek clarification of the lawyers' information request in as much detail and as often as they desired, and the lawyers were encouraged to continue requesting information from the database until they were satisfied they had enough information to defend the lawsuit on that particular issue or query. In the test, each query required a number of revisions, and the lawyers were not generally satisfied until many retrieved sets of documents had been generated and evaluated.

Precision was calculated by dividing the total number of relevant (i.e., "vital," "satisfactory," and "marginally relevant") documents retrieved by the total number of retrieved documents. If two or more retrieved sets were generated before the lawyer was satisfied with the results of the search, then the retrieved set considered for calculating Precision was computed as the *union* of all retrieved sets generated for that request. (Documents that appeared in more than one retrieved set were automatically excluded from all but one set.)

Recall was considerably more difficult to calculate since it required finding relevant documents that had not been retrieved in the course of the lawyers' search. To find the *unretrieved* relevant documents, we developed sample frames consisting of subsets of the unretrieved database that we believed to be rich in relevant documents (and from which duplicates of retrieved rel-

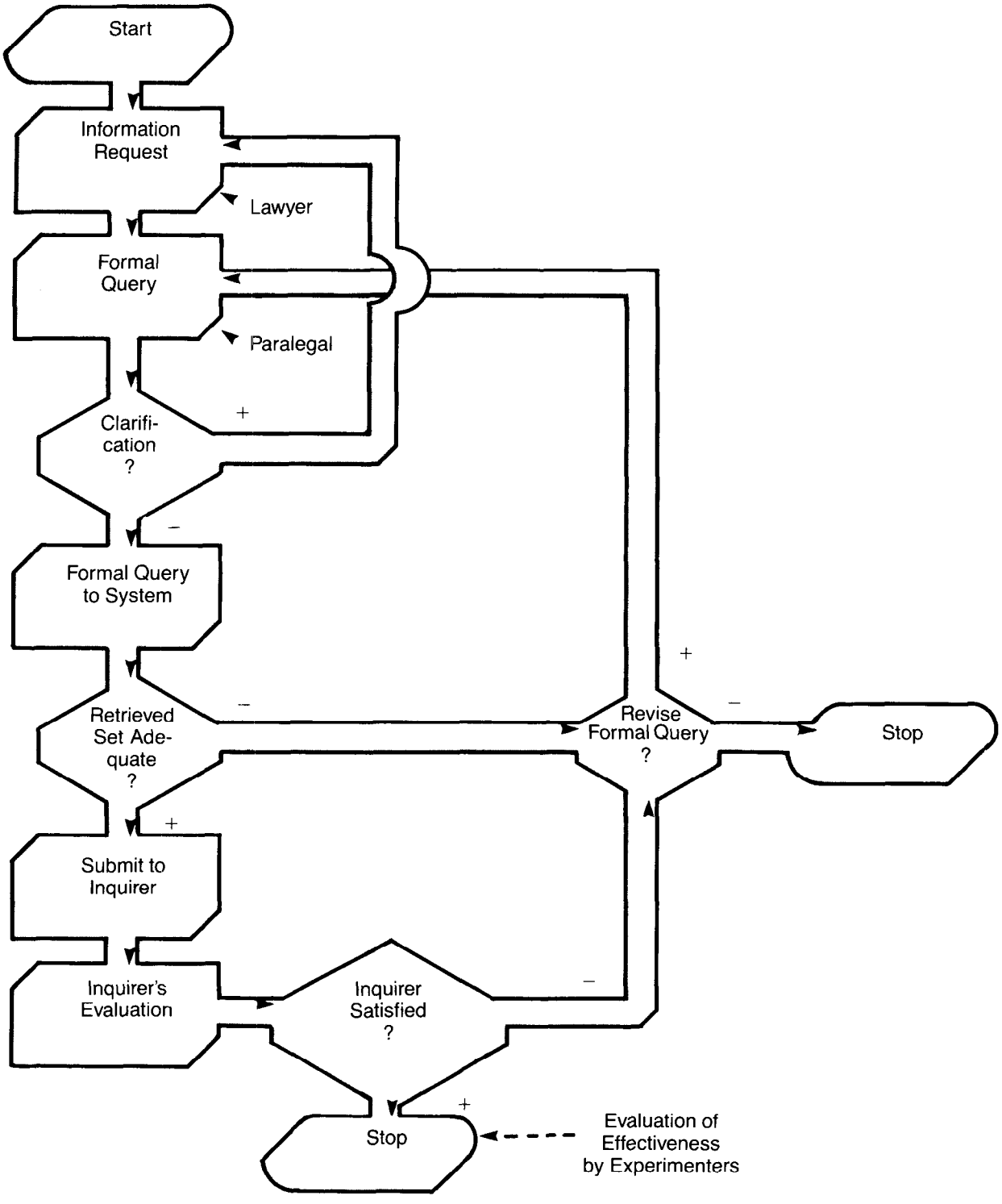

FIGURE 3. The Information Request Procedure

evant documents had been excluded). Random samples were taken from these subsets, and the samples were examined by the lawyers in a blind evaluation; the lawyers were not aware they were evaluating sample sets rather than retrieved sets they had personally generated. The total number of relevant documents that existed in these subsets could then be estimated. We sampled from subsets of the database rather than the entire database because, for most queries, the percentage of relevant documents in the database was less than

2 percent, making it almost impossible to have both manageable sample sizes and a high level of confidence in the resulting Recall estimates. Of course, no extrapolation to the entire database could be made from these Recall calculations. Nonetheless, the estimation of the number of relevant unretrieved documents in the subsets did give us a *maximum* value for Recall for each request.

TEST RESULTS

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

In Table I we show the values of Precision and Recall for each of the 40 requests. The values of Precision ranged from a maximum of 100.0 percent to a minimum of 19.6 percent. The unweighted average value of Precision turned out to be 79.0 percent (standard deviation = 23.2). The weighted average was 75.5 percent. This meant that, on average, 79 out of every 100 documents retrieved using STAIRS were judged to be relevant.

The values of Recall ranged from a maximum of 78.7 percent to a minimum of 2.8 percent. The unweighted average value of Recall was 20 percent (standard deviation = 15.9), and the weighted average value was 20.26 percent. This meant that, on average, STAIRS could be used to retrieve only 20 percent of the relevant documents, whereas the lawyers using the system believed they were retrieving a much higher percentage (i.e., over 75 percent).

When we plot the value of Precision against the corresponding value of Recall for each of the 40 information requests, we get the scatter diagram given in Figure 4. Although Figure 4 contains no more data than Table I, it does show the relationships in a more explicit way. For example, the heavy clustering of points in the lower right corner shows that in over 50 percent of the cases we get values of Precision above 80 percent with Recall at or below 20 percent. The clustering in the lower portion of the diagram shows that in 80 percent of the information requests the value of Recall was at or below 20 percent. Figure 4 also depicts the frequently observed inverse relationship between Recall and Precision, where high values of Precision are often accompanied by low values for Recall, and vice versa [8].

OTHER FINDINGS

After the initial Recall/Precision estimations were done, several other statistical calculations were carried out in the hope that additional inferences could be made. First, the results were broken down by lawyer to ascertain whether certain individuals were prima facie

TABLE I. Recall and Precision Values for Each Information Request

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

Average Recall = 20.0% ←(Standard deviation = 15.9)
Average Precision = 79.0% ←(Standard deviation = 23.3)

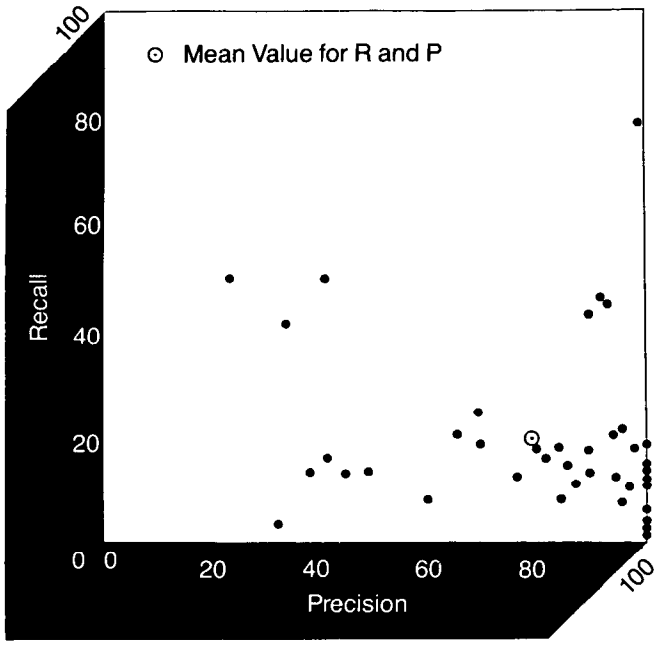

FIGURE 4. Plot of Precision versus Recall for All Information Requests

more adept at using the system than others. The results were as follows:

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

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

Another area of interest related to the revisions made to requests when the lawyer was not completely satisfied with the initial retrieved sets of documents. We hypothesized that if the values of Recall and Precision for the requests where substantial revisions had to be made (about 30 percent of the total) were significantly different from the overall mean values we might be able to infer something about the requesting procedure. Unfortunately, the values for Recall and Precision for the substantially revised queries (23.9 percent and 62.1 percent, respectively) did not indicate a statistically significant difference.

Finally, we tested the hypothesis that extremely high values of Precision for the retrieved sets would correlate directly with the lawyers' judgments of satisfaction with that set of documents (which might indicate that the lawyers were confusing Precision with Recall). To do this, we computed the mean Precision for all requests where the lawyers were satisfied with the initial retrieved set, and compared this value to the mean Precision for all requests. Although the Precision for requests that were not revised came out to be 85.4 percent, again the results were not statistically significant at the .05 level.

The Retrieval Effectiveness of Lawyers versus Paralegals

The argument can be made that, because STAIRS is a high-speed, on-line, interactive system, the searcher at the terminal can quickly and effectively evaluate the output of STAIRS during the query modification process. Therefore, retrieval effectiveness might be significantly improved if the person originating the information request is actually doing the searching at the terminal. This would mean that if a lawyer worked directly on the query formulation and query modification at the STAIRS terminal, rather than using a paralegal as intermediary, retrieval effectiveness might be improved.

We tested this conjecture by comparing the retrieval effectiveness of the lawyer vis à vis the paralegal on the *same* information request. We selected (at random) five information requests for which the searches had already been completed by the paralegal, and for which retrieved sets had been evaluated by the lawyer and values of Recall computed. (Neither the lawyer who made the relevance judgments nor the paralegal knew the Recall figures for these original requests.) We invited the lawyer to use STAIRS directly to access the database, giving the lawyer copies of his or her original information requests. The lawyer translated these requests into formal queries, evaluating the text displayed on the screen, modifying the queries as he or she saw fit, and finally deciding when to terminate the search. For each of the five information requests, we estimated the minimum number of relevant documents in the entire file, and knowing which documents the lawyer had previously judged relevant, we were able to compute the values of Recall for the lawyer at the terminal as we had already done for the paralegal. If it were true that STAIRS would give better results when the lawyers themselves worked at the terminal, the values of Recall for the lawyers would have to be significantly higher than the values of Recall when the paralegals did the searching. The results were as follows:

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

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

both the lawyer and the paralegal get the same results for Recall.

WHY WAS RECALL SO LOW

The realization that STAIRS may be retrieving only one out of five relevant documents in response to an information request may surprise those who have used STAIRS or had it demonstrated to them. This is because they will have seen only the retrieved set of documents and not the total corpus of relevant documents; that is, they have seen that the proportion of relevant documents in the retrieved set (i.e., Precision) is quite good (around 80 percent). The important issues to consider here are (1) why was Recall so low and (2) why did the users (lawyers and paralegals) believe they were retrieving 75 percent of the relevant documents when, in fact, they were only retrieving 20 percent.

The low values of Recall occurred because full-text retrieval is difficult to use to retrieve documents by subject because its design is based on the assumption that it is a simple matter for users to foresee the exact words and phrases that will be used in the documents they will find useful, and *only* in those documents. This assumption is not a new one; it goes back over 25 years to the early days of computing. The basic idea is that one can use the formal aspects of text to predict its meaning or subject content: formal aspects such as the occurrence, location, and frequency of words; and to the extent that it can be precisely described, the syntactic structure of word phrases. It was hoped that by exploiting the high speed of a computer to analyze the formal aspects of text, one could get the computer to deal with text in a "comprehending-like" way (i.e., to identify the subject content of texts). This endeavor is known as "Automatic Indexing" or, in a more general sense, "Natural Language Processing." During the past two decades, many experiments in automatic indexing (of which full-text searching is the simplest form) have been carried out, and many discussions by linguists, psychologists, philosophers, and computer scientists have analyzed the results and the issues [5]. These experiments show that full-text document retrieval has worked well only on unrealistically small databases.

The belief in the predictability of the words and phrases that may be used to discuss a particular subject is a difficult prejudice to overcome. In a naive sort of way, it is an appealing prejudice but a prejudice nonetheless, because the effectiveness of full-text retrieval has not been substantiated by reliable Recall measures on realistically large databases. Stated succinctly, it is impossibly difficult for users to predict the exact words, word combinations, and phrases that are used by *all* (or most) relevant documents and *only* (or primarily) by those documents, as can be seen in the following examples.

In the legal case in question, one concern of the lawyers was an accident that had occurred and was now an object of litigation. The lawyers wanted all the reports, correspondence, memoranda, and minutes of meetings that discussed this accident. Formal queries were constructed that contained the word "accident(s)" along with several relevant proper nouns. In our search for *unretrieved* relevant documents, we later found that the accident was not always referred to as an "accident," but as an "event," "incident," "situation," "problem," or "difficulty," often without mentioning any of the relevant proper names. The manner in which an individual referred to the incident was frequently dependent on his or her point of view. Those who discussed the event in a critical or accusatory way referred to it quite directly—as an "accident." Those who were personally involved in the event, and perhaps culpable, tended to refer to it euphemistically as, inter alia, an "unfortunate situation," or a "difficulty." Sometimes the accident was referred to obliquely as "the subject of your last letter," "what happened last week was . . . ," or, as in the opening lines of the minutes of a meeting on the issue, "Mr. A: We all know why we're here . . . ." Sometimes relevant documents dealt with the problem by mentioning only the technical aspects of why the accident occurred, but neither the accident itself nor the people involved. Finally, much relevant information discussed the situation *prior* to the accident and, naturally, contained no reference to the accident itself.

Another information request resulted in the identification of 3 key terms or phrases that were used to retrieve relevant information; later, we were able to find 26 other words and phrases that retrieved additional relevant documents. The 3 original key terms could not have been used individually as they would have retrieved 420 documents, or approximately 4000 pages of hard copy, an unreasonably large set, most of which contained irrelevant information. Another request identified 4 key terms/phrases that retrieved relevant documents, which we were later able to enlarge by 44 additional terms and combinations of terms to retrieve relevant documents that had been missed.

Sometimes we followed a trail of linguistic creativity through the database. In searching for documents discussing "trap correction" (one of the key phrases), we discovered that relevant, unretrieved documents had discussed the same issue but referred to it as the "wire warp." Continuing our search, we found that in still other documents trap correction was referred to in a third and novel way: the "shunt correction system." Finally, we discovered the inventor of this system was a man named "Coxwell" which directed us to some documents he had authored, only he referred to the system as the "Roman circle method." Using the Roman circle method in a query directed us to still more relevant but unretrieved documents, but this was not the end either. Further searching revealed that the system had been tested in another city, and all documents germane to those tests referred to the system as the "air truck." At this point the search ended, having consumed over an entire 40-hour week of on-line searching, but there is no reason to believe that we had reached the end of the trail; we simply ran out of time.

As the database included many items of personal cor-

respondence as well as the verbatim minutes of meetings, the use of slang frequently changed the way in which one would "normally" talk about a subject. Disabled or malfunctioning mechanisms with which the lawsuit was concerned were sometimes referred to as "sick" or "dead," and a burned-out circuit was referred to as being "fried." A critical issue was sometimes referred to as the "smoking gun."

Even misspellings proved an obstacle. Key search terms like "flattening," "gauge," "memos," and "correspondence," which were essential parts of phrases, were used effectively to retrieve relevant documents. However, the misspellings "flatening," "guage," "gage," "memoes," and "correspondance," using the same phrases, also retrieved relevant documents. Misspellings like these, which are tolerable in normal everyday correspondence, when included in a computerized database become literal traps for users who are asked not only to anticipate the key words and phrases that may be used to discuss an issue but also to foresee the whole range of possible misspellings, letter transpositions, and typographical errors that are likely to be committed.

Some information requests placed almost impossible demands on the ingenuity of the individual constructing the query. In one situation, the lawyer wanted "Company A's comments concerning . . . ." Looking at the documents authored by Company A was not enough, as many relevant comments were embedded in the minutes of meetings or recorded secondhand in the documents authored by others. Retrieving all the documents in which Company A was mentioned was too broad a search; it retrieved over 5,000 documents (about 40,000+ pages of hard copy). However, predicting the exact phraseology of the text in which Company A commented on the issue was almost impossible; sometimes Company A was not even mentioned, only that so-and-so (representing Company A) "said/considered/remarked/pointed out/commented/noted/explained/discussed," etc.

In some requests, the most important terms and phrases were not used at all in relevant documents. For example, "steel quantity" was a key phrase used to retrieve important relevant documents germane to an actionable issue, but unretrieved relevant documents were also found that did not report *steel quantity* at all, but merely the *number* of such things as "girders," "beams," "frames," "bracings," etc. In another request, it was important to find documents that discussed "nonexpendable components." In this case, relevant unretrieved documents merely listed the names of the components (of which there were hundreds) and made no mention of the broader generic description of these items as "nonexpendable."

Why didn't the lawyers realize they were not getting all of the information relevant to a particular issue? Certainly they knew the lawsuit. They had been involved with it from the beginning and were the principal attorneys representing the defense. In addition, one of the paralegals had been instrumental not only in setting up the database but also in supervising the selection of relevant information to be put on-line. Might it not be reasonable to expect them to be suspicious that they were not retrieving everything they wanted? Not really. Because the database was so large (providing access to over 350,000 pages of hard copy, all of which was in some way pertinent to the lawsuit), it would be unreasonable to expect four individuals (two lawyers and two paralegals) to have total recall of all the important supporting facts, testimony, and related data that were germane to the case. If they had such recall they would have no need for a computerized, interactive retrieval system. It is well known among cognitive psychologists that man's power of literal recall is much less effective than his power of recognition. The lawyers could remember the exact text of some of the important information, but as we have already stated, this was a very small subset of the total information relevant to a particular issue. They could *recognize* the important information when they saw it, and they could do so with uncanny consistency. (As a control, we submitted some retrieved sets and sample sets of documents to the lawyers several times in a blind test of their evaluation consistency, and found that their consistency was almost perfect.) Also, since the lawyers were not experts in information retrieval system design, there were no a priori reasons for them to suspect the Recall levels of STAIRS.

DETERIORATION OF RECALL AS A FUNCTION OF FILE SIZE

One reason why Recall evaluations done on small databases cannot be used to estimate Recall on larger databases is because, ceteris paribus, the value of Recall decreases as the size of the database increases, or, from a different point of view, the amount of search effort required to obtain the same Recall level increases as the database increases, often at a faster rate than the increase in database size. On the database we studied, there were many search terms that, used by themselves, would retrieve over 10,000 documents. Such *output overload* is a frequent problem of full-text retrieval systems.

As a retrieved set of several thousand documents is impractical, the user must reduce the output overload by reformulating the single-term query so that it retrieves fewer documents. If a single term query w_1 retrieves too many documents, the user may add another term, w_2, so as to form the new query "w_1 and w_2" (or "w_1 adjacent w_2," or "w_1 same w_2"). The reformulated query cannot retrieve more documents than the original; most probably, it will retrieve many fewer. The process of adding intersecting terms to a query can be continued until the size of the output reaches a manageable number. (This strategy, and its consequences, is discussed in more detail in [1].) However, as the user narrows the size of the output by adding intersecting terms, the value of Recall goes down because, with each new term, the probability is that some relevant documents will be excluded by that reformulated query.

The deterioration of Recall from a probabilistic point of view is quite startling. For each query, there is a class of relevant documents that we designate as R. We represent the probability that each of those documents will contain some word w_1 as p, and the probability that a relevant document will contain some other word w_2 as q. Thus, the value of Recall for a request using only w_1 will be equal to p, and Recall for a request using only w_2 will be equal to q. Now the probability that a relevant document will contain both w_1 and w_2 is less than or equal to either p or q. If we assume that the respective appearances of w_1 and w_2 in a relevant document are independent events, then the probability of both of them appearing in a relevant document would be equal to the product of p and q. Since both p and q are usually numbers less than unity, their product usually will be smaller than either p or q. This means that Recall, which can also be thought of as the probability of retrieving a relevant document, is now equal to the product of p and q. In other words, reducing the number of documents retrieved by intersecting an increasing number of terms in the formal query causes Recall for that query also to decrease.

However, the problem is really much worse. In order for a relevant document, which contains w_1 and w_2, to be retrieved by a single query, a searcher must select and use those words in his or her query. The probability that the searcher will select w_1 is, of course, generally less than 1.0; and the probability that w_1 will occur in a relevant document is also usually less than 1.0. However, these probabilities must be multiplied by the probability that the searcher will select w_2 as part of his or her query, and the probability that w_2 will occur in a relevant document. Thus, calculating Recall for a two-term search involves the multiplication of four numbers each of which is usually less than 1.0. As a result, the value of Recall gets very small (see Table II). When we consider a three- or four-term query, the value of Recall drops off even more sharply.

TABLE II. The Probability of Retrieving a Relevant Document Containing Terms w_1 and w_2

$P(Sw_1) = .6$ = Probability searcher uses term w_1 in a search query

$P(Sw_2) = .5$ = Probability searcher uses term w_2 in a search query

$P(Dw_1) = .7$ = Probability w_1 appears in a relevant document

$P(Dw_2) = .6$ = Probability w_2 appears in a relevant document

Probability of searcher selecting w_1 and a relevant document containing w_1:

$$P(Sw_1) \times P(Dw_1) = (.6) \times (.7) = .42$$

Probability of searcher selecting w_2 and a relevant document containing w_2:

$$P(Sw_2) \times P(Dw_2) = (.5) \times (.6) = .30$$

Probability of searcher selecting w_1 and w_2 and a relevant document containing w_1 and w_2:

$$P(Sw_1) \times P(Dw_1) \times P(Sw_2) \times P(Dw_2)$$

$$\text{(e.g., } P(.6) \times P(.7) \times P(.5) \times P(.6) = .126$$

The problem of output overload is especially critical in full-text retrieval systems like STAIRS, where the frequency of occurrence of search terms is considerably larger than (and increases faster than) the frequency of occurrence (or "breadth") of index terms in a database where the terms are manually assigned to documents. This means that the user of a full-text retrieval system will face the problem of output overload sooner than the user of a manually indexed system. The solution that STAIRS offers—conjunctively adding search terms to the query—does reduce the number of documents retrieved to a manageable number but also eliminates relevant documents. Search queries employing four or five intersecting terms were not uncommon among the queries used in our test. However, the probability that a query that intersects five terms will retrieve relevant documents is quite small. If we were to assign a probability of .7 to all the respective probabilities in a hypothetical five-term query as we did in the two-term query in Table II (and .7 is an optimistic average value), the Recall level for that query would be .028. In other words, that query could be expected to retrieve *less than* 3 percent of the relevant documents in the database. If the probabilities for the five-term query were a more realistic average of .5, the Recall value for that query would be .0009! This means that if there were 1000 relevant documents on the database, it is likely that this query would retrieve only one of them. The searcher must submit many such low-yield queries to the system if he or she wants to retrieve a high percentage of the relevant documents.

DISCUSSION

The reader who is surprised at the results of this test of retrieval effectiveness is not alone. The lawyers who participated in the test were equally astonished. Although there are sound theoretical reasons why we should expect these results, they seem to run counter to previous tests of retrieval effectiveness for full-text retrieval.

Two pioneering evaluations of full-text retrieval systems by respected researchers in the field (Swanson [6] and Salton [3]) determined to their satisfaction that full-text document-retrieval systems could retrieve relevant documents at a satisfactory level while avoiding the problems of manual indexing. Our study, on the other hand, shows that full-text document retrieval does *not* operate at satisfactory levels and that there are sound theoretical reasons to expect this to be so. Who is right? Well, we all are, and this is not an equivocation. The two earlier studies drew the correct conclusions from their evaluations, but these conclusions were different from ours because they were based on small experimental databases of less than 750 documents. Our study was done not on an experimental database but an actual, operational database of almost 40,000 documents. Had Swanson and Salton been fortunate enough to study a retrieval system as large as ours, they

would undoubtedly have observed similar phenomena (Swanson was later to comment perceptively on the difficulty of drawing accurate conclusions about document retrieval from experiments using small databases [7]). In addition, it has only recently been observed that information-retrieval systems do not scale up [2]. That is, retrieval strategies that work well on small systems do not necessarily work well on larger systems (primarily because of output overload). This means that studies of retrieval effectiveness must be done on full-sized retrieval systems if the results are to be indicative of how a large, operational system would perform. However, large-scale, detailed retrieval-effectiveness studies, like the one reported here, are unprecedented because they are incredibly expensive and time consuming; our experiment took six months; involved two researchers and six support staff; and, taking into account all direct and indirect expenses, cost almost half a million dollars. Nevertheless, Swanson and Salton's earlier full-text evaluations remain pioneering studies and, rather than contradict our findings, have an illuminating value of their own.

An objection that might be made to our evaluation of STAIRS is that the low Recall observed was not due to STAIRS but rather to query-formulation error. This objection is based on the realization that, at least in principle, virtually any subset of the database is retrievable by some simple or complex combination of search terms. The user's task is simply to find the right combination of search terms to retrieve *all* and *only* the relevant documents. However, we believe that users should not be asked to shoulder the blame, and perhaps an analogy will indicate why. Suppose you ask a company to make a lock for you, and they oblige by providing a combination lock; but when you ask them for the combination to open the lock, they say that finding the correct combination is your problem, not theirs. Now, it is possible, in principle, to find the correct combination, but in practice it may be impossibly difficult to do so. A full-text retrieval system bears the burden of retrieval failure because it places the user in the position of having to find (in a relatively short time) an impossibly difficult combination of search terms. The person using a full-text retrieval system to find information on a relatively large database is in the same unenviable position as the individual looking for the combination to the lock. It is true that we, as evaluators, found the combinations of search terms necessary to retrieve many of the unretrieved relevant documents, but three things should be kept in mind. First, we make no claim to having found *all* the relevant unretrieved documents; we may not have found even half of them, as our sampling technique covered only a small percentage of the database. Second, a tremendous amount of search time was involved with *each* request (sometimes over 40 hours of on-line time), and the entire test took almost 6 months. Such inefficiency is clearly not consonant with the high speed desired for computerized retrieval. Third, the evaluators in this case represented, together, over 40 years of practical and theoretical experience in information systems analysis and should be expected to have somewhat better searching abilities than the typical STAIRS searcher. Moreover, STAIRS is sold under the premise that it is easy to use and requires no sophisticated training on the part of the user. Yet this study is a clear demonstration of just how sophisticated search skills must be to use STAIRS, or, mutatis mutandis, any other full-text retrieval system. There is evidence that this problem is beginning to be recognized by at least one full-text retrieval vendor, WESTLAW, which has made its reputation by offering full-text access to legal cases. WESTLAW has now begun to supplement its full-text retrieval with manually assigned index terms.

SUMMARY

This paper has presented a major, detailed evaluation of a full-text document-retrieval system. We have shown that the system did not work well in the environment in which it was tested and that there are theoretical reasons why full-text retrieval systems applied to large databases are unlikely to perform well in any retrieval environment. The optimism of early studies was based on the small size of the databases used, and were geared toward showing only that full-text search was *competitive* with searching based on manually assigned index terms, under the assumption that, if it were competitive, full-text retrieval would eliminate the cost of indexing. However, there are costs associated with a full-text system that a manual system does not incur. First, there is the increased time and cost of entering the full text of a document rather than a set of manually assigned subject and context descriptors. The average length of a document record on the system we evaluated was about 10,000 characters. In a manually assigned index-term system of the same type, we found the average document record to be less than 500 characters. Thus, the full-text system incurs the additional cost of inputting and verifying *20 times* the amount of information that a manually indexed system would need to deal with. This difference alone would more than compensate for the added time needed for manual indexing and vocabulary construction. The 20-fold increase in document record size also means that the database for a full-text system will be some 20 times larger than a manually indexed database and entail increased storage and searching costs. Finally, because the average number of searchable subject terms per document for the full-text retrieval system described here was approximately 500, whereas a manually indexed system might have a subject indexing depth of about 10, the dictionary that lists and keeps track of these assignments (i.e., provides pointers to the database) could be as much as *50 times* larger on a full-text system than on a manually indexed system. A full-text retrieval system does not give us something for nothing. Full-text searching is one of those things, as Samuel Johnson put it so succinctly, that "· · · is never done well, and one is surprised to see it done at all."

Acknowledgments. The authors would like to thank William Cooper of the University of California at Berkeley for his comments on an earlier version of this manuscript, and Barbara Blair for making the drawings that accompany the text.

REFERENCES

1. Blair, D.C. Searching biases in large interactive document retrieval systems. *J. Am. Soc. Inf. Sci. 31* (July 1980), 271–277.
2. Resnikoff, H.L. The national need for research in information science. STI Issues and Options Workshop, House subcommittee on science, research and technology, Washington, D.C., Nov. 3, 1978.
3. Salton, G. Automatic text analysis. *Science 168*, 3929 (Apr. 1970), 335–343.
4. Saracevic, T. Relevance: A review of and a framework for thinking on the notion in information science. *J. Am. Soc. Inf. Sci. 26* (1975), 321–343.
5. Sparck Jones, K. *Automatic Keyword Classification for Information Retrieval.* Butterworths, London, 1971.
6. Swanson, D.G. Searching natural language text by computer. *Science 132*, 3434 (Oct. 1960), 1099–1104.
7. Swanson, D.R. Information retrieval as a trial and error process. *Libr. Q. 47*, 2 (1978), 128–148.
8. Swets, J.A. Information retrieval systems. *Science 141* (1963), 245–250.
9. Zunde, P., and Dexter, M.E. Indexing consistency and quality. *Am. Doc. 20*, 3 (July 1969), 259–264.

CR Categories and Subject Descriptors: H.1.0 [**Models and Principles**]: General; H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval—*search process, query formulation*
General Terms: Design, Human Factors, Theory
Additional Key Words and Phrases: full-text document retrieval, litigation support, retrieval evaluation, Recall and Precison

Received 4/84; accepted 9/84

Authors' Present Addresses: David C. Blair, Graduate School of Business Administration, The University of Michigan, Ann Arbor, MI 48109; M.E. Maron, School of Library and Information Studies, The University of California, Berkeley, CA 94720.