Division of Research
Graduate School of Business Administration
The University of Michigan

February 1984

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

Working Paper No. 364

David C. Blair
The University of Michigan

and

M.E. Maron
University of California, Berkeley

Abstract

A detailed evaluation of a large, operational full-text document retrieval system is described. Values of Precision and Recall are estimated using statistical sampling methods and blind evaluation procedures. The results of the evaluation show that the system tested was retrieving less than 20% of the relevant documents when the searchers believed it was retrieving over 75% of the relevant documents. These findings are discussed and are explained in terms of the theory and assumptions of full-text document retrieval.

Introduction

Document retrieval is everyone's problem. It is the problem of how to find the stored documents that contain useful information. Here is a more precise formulation of the document retrieval problem: There exists 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. In addition to this set of documents there exist individuals who, at different times and for different reasons, are looking for recorded information--information that may be contained in some of the documents of this set. For each instance in which an individual seeks information, he will find some of the documents of that set useful and other documents not useful. Those documents that he finds useful are, we say, relevant for him; the others are not relevant.

How should a collection of documents be organized so that a person in search of useful information can find all and only the relevant items? One proposal for solving this problem is to use automatic full-text retrieval. The basic notion 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, he "merely" instructs the computer to search for all those documents which contain certain words and word combinations which he has specified.

Two elements make this idea of automatic full-text retrieval even more attractive: On the one hand, digital technology continues to provide computers which are larger, faster, cheaper, more reliable and easier to use thereby making large scale full-text retrieval increasingly feasible. On the other

hand, full-text retrieval entails the elimination of human indexers who are increasingly costly to employ and whose indexing 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 [1]. He 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 [2].

This paper describes a large scale full-text search and retrieval experiment aimed at evaluating the effectiveness of full-text retrieval. For our study we used IBM's full-text retrieval system, STAIRS. STAIRS, which is an acronym for "STorage and Information Retrieval System," is a very fast, large capacity, full-text, document retrieval system. Our empirical study of STAIRS in a litigation support system situation, showed that its retrieval effectiveness was surprisingly poor. We offer theoretical reasons to explain why this poor performance should not be surprising, and we explain 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 full-text retrieval system; therefore, our study should not be construed as a critique of STAIRS alone, but a critique of the "principles" on which it and other full-text document retrieval systems are based.

The Allure of Full-Text Document Retrieval

The retrieval of texts of documents by subject content occupies a special place in the province of information retrieval for, unlike data retrieval, the richness and flexibility of natural language have a significant influence

on the conduct of an inquirer's search. The inquirer must describe his information need using subject descriptors actually assigned to documents on the data base he is searching, while the indexer must choose appropriate subject terms to describe the "information content" of the documents to be included in that database (see Figure 1). But there are no clear and precise rules which an indexer can follow to select the "appropriate subject terms" describing a particular document. This means that even trained indexers may be inconsistent in the selection of subject terms to describe documents. Experimental studies on indexing have confirmed this by demonstrating that different indexers will generally index the same document differently [3]. (Even the same individual will not always select the same index terms if asked at a later time to index a document he has previously indexed.) Such problems with the manual assignment of subject descriptors to documents makes computerized, full-text document retrieval appealing. By entering the entire, or the most significant part of, the text of a document onto the database one is freed, it is argued, from the inherent evils of manually creating document records which reflect the subject content of a particular document. The evils avoided include: the construction of an indexing vocabulary, the training of indexers, the excessive time needed to scan/read the documents and assign context and subject terms to documents. Such economies are appealing, but for full-text retrieval to be worthwhile it must also provide satisfactory levels of retrieval effectiveness. The following experiment was conducted in order to measure accurately these 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

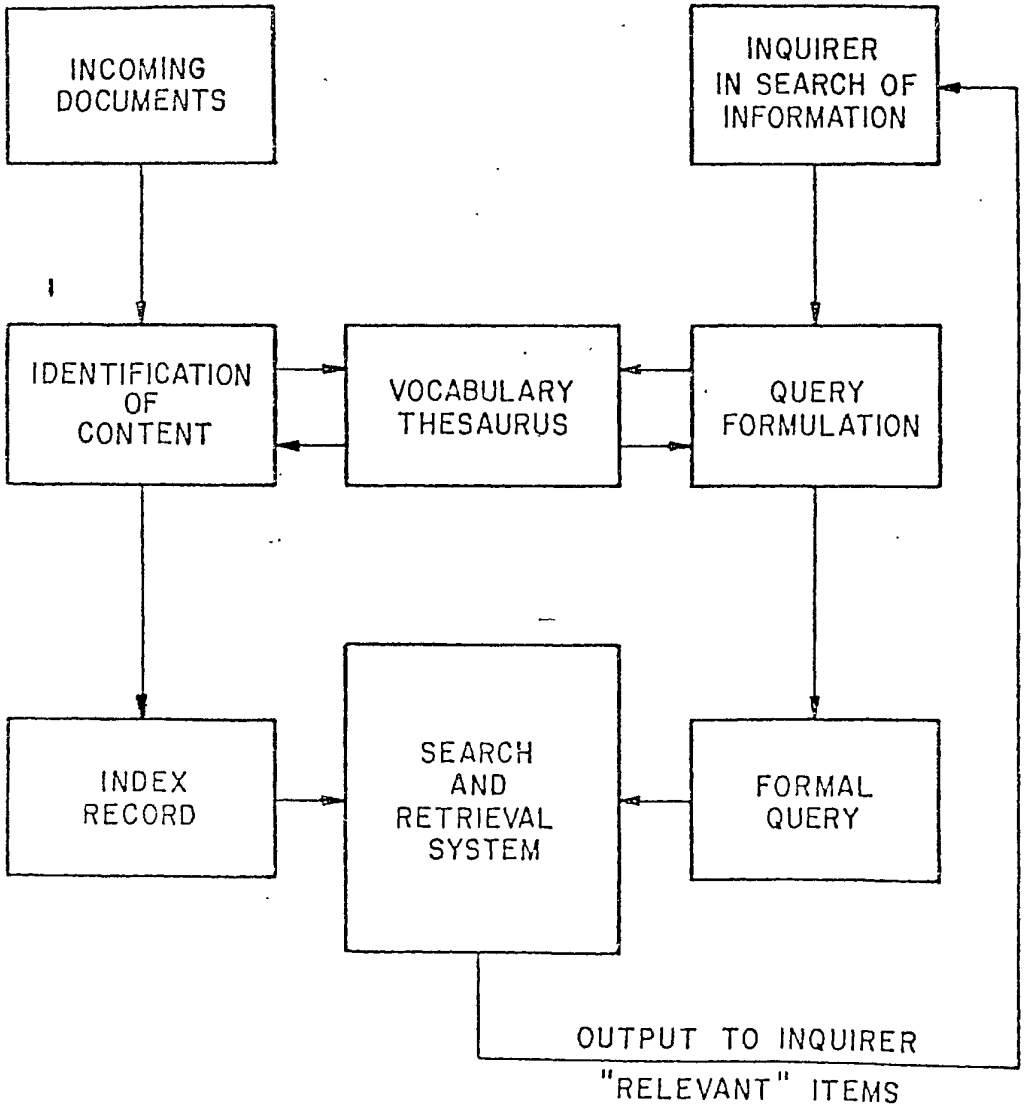
INCOMING DOCUMENTS
INQUIRER IN SEARCH OF INFORMATION
IDENTIFICATION OF CONTENT
VOCABULARY THESAURUS
QUERY FORMULATION
INDEX RECORD
SEARCH AND RETRIEVAL SYSTEM
FORMAL QUERY
OUTPUT TO INQUIRER
"RELEVANT" ITEMS

Figure 1

of the relevant documents and Precision measures how well the system retrieves only the relevant documents (For the purpose of this study, we understand "relevance" to be something quite straight-forward: a document is relevant if it is judged useful by the inquirer who initiated the search; if not, then it is non-relevant, see [4].). In order to be more precise about the definitions of Recall and Precision, look at Figure 2. Recall is the proportion of relevant documents which the system retrieves. This means that Recall is the ratio of $\frac{x}{n_2}$. 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 $\frac{x}{n_1}$, 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 approximately 350,000 pages of hard-copy text. These documents were for use in the defense of a large corporate law suit, and access to the information was provided by IBM's STAIRS/TLS software (STorage And Information Retrieval System/Thesaurus Linguistic System). STAIRS software represents state-of-the-art software in full-text retrieval. It provides facilities for retrieving text where specified words appear singly or in complex Boolean combinations. An inquirer can specify the retrieval of text where words appear together anywhere in the document, within the same paragraph, within the same sentence, or adjacent to each other (as in "New" adjacent "York"). Retrieval can also be performed on fields other than the text of the document, such as: author; date; and, document number. STAIRS also provides ranking functions which could be used to order retrieved sets

$$\text{RECALL} = \frac{\text{NUMBER OF RELEVANT AND RETRIEVED}}{\text{TOTAL NUMBER RELEVANT}} = \frac{x}{n_2}$$

$$\text{PRECISION} = \frac{\text{NUMBER OF RELEVANT AND RETRIEVED}}{\text{TOTAL NUMBER RETRIEVED}} = \frac{x}{n_1}$$

Figure 2

of 200 or less documents. These functions permit the inquirer to order retrieved sets in ascending or descending numerical (e.g., dates) or alphabetic (e.g., authors) order. In addition, retrieved sets of less than 200 documents could be ordered by the frequency in which specified search terms occurred in the retrieved documents. The Thesaurus Linguistic System provides the facilities to manually create an interactive thesaurus which could be called by an inquirer to semantically broaden his searches. The TLS provides the tools for the designer to specify such semantic relationships between search terms as "narrower than," "broader than," "related to," "synonomous with," and automatic phrase decomposition. STAIRS/TLS thus represents a complete full-text document retrieval system.

The Experimental Protocol

We wanted to test how well STAIRS could be used to retrieve all and only the documents relevant to a given request for information. In essence, we wanted to determine the values of Recall (percentage of relevant documents retrieved), and Precision (percentage of retrieved documents that are relevant). While Precision is an important measure of retrieval effectiveness, it is meaningless unless compared to the level of Recall desired by the inquirers. In this case, the lawyers who were to use the system for litigation support stipulated that they must be able to retrieve at least 75% of all the documents relevant to a given request for information, and 100% of those documents they regarded as "vital" to the defense of the case. (The lawyers, as was their custom, 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 manner it would have been used during actual litigation. Two lawyers, who were the principal defense attorneys in the suit, participated in the experiment. They generated a total of 51 different information requests, and these requests were translated into formal queries by either of two paralegals, both of whom were familiar with the case and experienced with the STAIRS retrieval system. The paralegals would search on the database until they found a set of documents which they believed would satisfy one of the initial requests made by the lawyers. The original hard copies of these documents were retrieved from files, and xerox copies of them were sent to the lawyer who originated the request. The lawyer would then evaluate the retrieved documents ranking them according to whether they were "vital," "satisfactory," "marginally relevant," or "irrelevant" to their original information request. The lawyer would then make an overall judgment concerning the retrieved set he had received, stating whether he wanted further refinement of the query and further searching for relevant documents. His reasons for any subsequent query revisions were made in writing and were fully recorded. The information request and query formulation procedures were considered to be complete only when the lawyer stated in writing that he was satisfied with the search results for that particular query (i.e., in his estimation he had more than 75% of the "vital," "satisfactory" and "marginally relevant" documents). It was only at this point that the experimenters could begin the task of measuring Precision and Recall. (A diagram of the information request procedure is given in Figure 3). It is important to emphasize that the lawyers and paralegals were permitted as much interaction as they thought necessary to insure highly effective retrieval. The paralegals could seek clarification of the

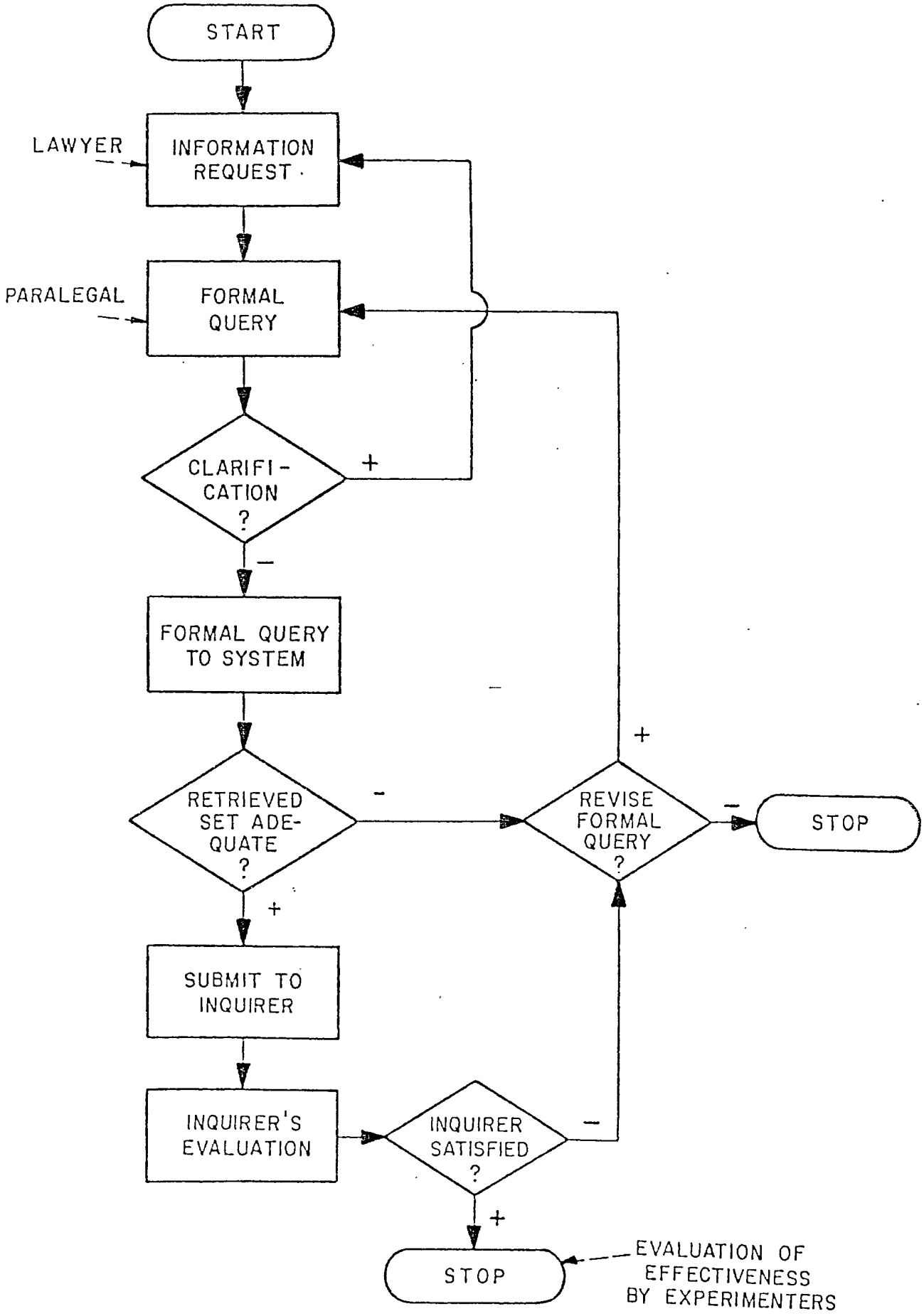
START
LAWYER
INFORMATION REQUEST
PARALEGAL
FORMAL QUERY
CLARIFI-CATION ?
+
-
FORMAL QUERY TO SYSTEM
RETRIEVED SET ADE-QUATE ?
-
+
REVISE FORMAL QUERY ?
+
-
STOP
SUBMIT TO INQUIRER
INQUIRER'S EVALUATION
INQUIRER SATISFIED ?
-
+
STOP
EVALUATION OF EFFECTIVENESS BY EXPERIMENTERS

Figure 3

lawyers' information request in as much detail and as often as they desired. The lawyers were encouraged to continue requesting information from the database until they were satisfied that they had enough information to defend the lawsuit on that particular issue (query). In the conduct of the experiment every query required a significant number of revisions, and the lawyers were not generally satisfied until many retrieved sets were generated and evaluated.

Precision was calculated by dividing the total number of relevant (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 information request (documents which appeared in more than one retrieved set were, of course, automatically excluded from all but one set).

Recall was considerably more difficult to calculate since it required us to find relevant documents that had not been retrieved in the course of the lawyers' searchers. To find these unretrieved relevant documents we developed sample frames consisting of subsets of the unretrieved database which we believed to be rich in relevant documents (from which duplicates of retrieved relevant documents had been excluded). Random samples were taken from these subsets and these samples were examined by the lawyers in a blind evaluation (i.e., the lawyers were not aware that they were evaluating sample sets rather than retrieved sets they had 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 from the entire database because, for most queries, the percentage of relevant documents in the database was less than 2%, making it almost impossible to have manageably small

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

Average Recall = 20.0%<——————(Standard Deviation = 15.9)

Average Precision = 79.0%<——————(Standard Deviation = 23.3)

Table 1

sample sizes and a high level of confidence. Of course, no extrapolation could be made to the entire database from these Recall calculations, but the estimation of the number of relevant unretrieved documents in these subsets of the database would give us a maximum value for Recall for each information request.

Text Results

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

In Table 1 we show the values of Precision and Recall for each of the 40 information requests mentioned above. The reader should remember that in making these calculations, a relevant document was any document judged by the lawyer as being either "vital," "satisfactory," or "marginally relevant." The values of Precision ranged from a maximum of 100% to a minimum of 19.6%. The unweighted average value of Precision turned out to be 79% (standard deviation = 23.2). The weighted average was 75.5%. This meant that, on the average, 79 out of every 100 documents retrieved using STAIRS were judged to be relevant.

The values of Recall ranged from a maximum of 78.7% to a minimum of 2.8%. The unweighted average value of Recall was 20% (standard deviation = 15.9). The weighted average value was 20.26%. This meant that, on the average, STAIRS could be used to retrieve only 20% of those documents that would be judged relevant when the inquirers believed that they were retrieving a much higher percentage of the relevant documents (the lawyers believed they were retrieving over 75% of the relevant documents at the time).

When we plot the value of Precision against the corresponding value of Recall for each of the 40 information requests, we get the scatter diagram shown in Figure 4. Although this scatter diagram does not contain any more data than is contained in Table 1, it does reveal the relationships in a more explicit manner. We can see, for example, a heavy clustering of points in the lower right corner. This shows that in over 50% of the cases we get values of Precision above 80% with Recall at or below 20%. Looking at the lower portion of the scatter diagram we see a clustering of points showing that in 80% of the information requests the value of Recall was at or below 20%. The diagram also depicts the well-known inverse relationship between Recall and Precision [5].

Other Findings

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

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

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

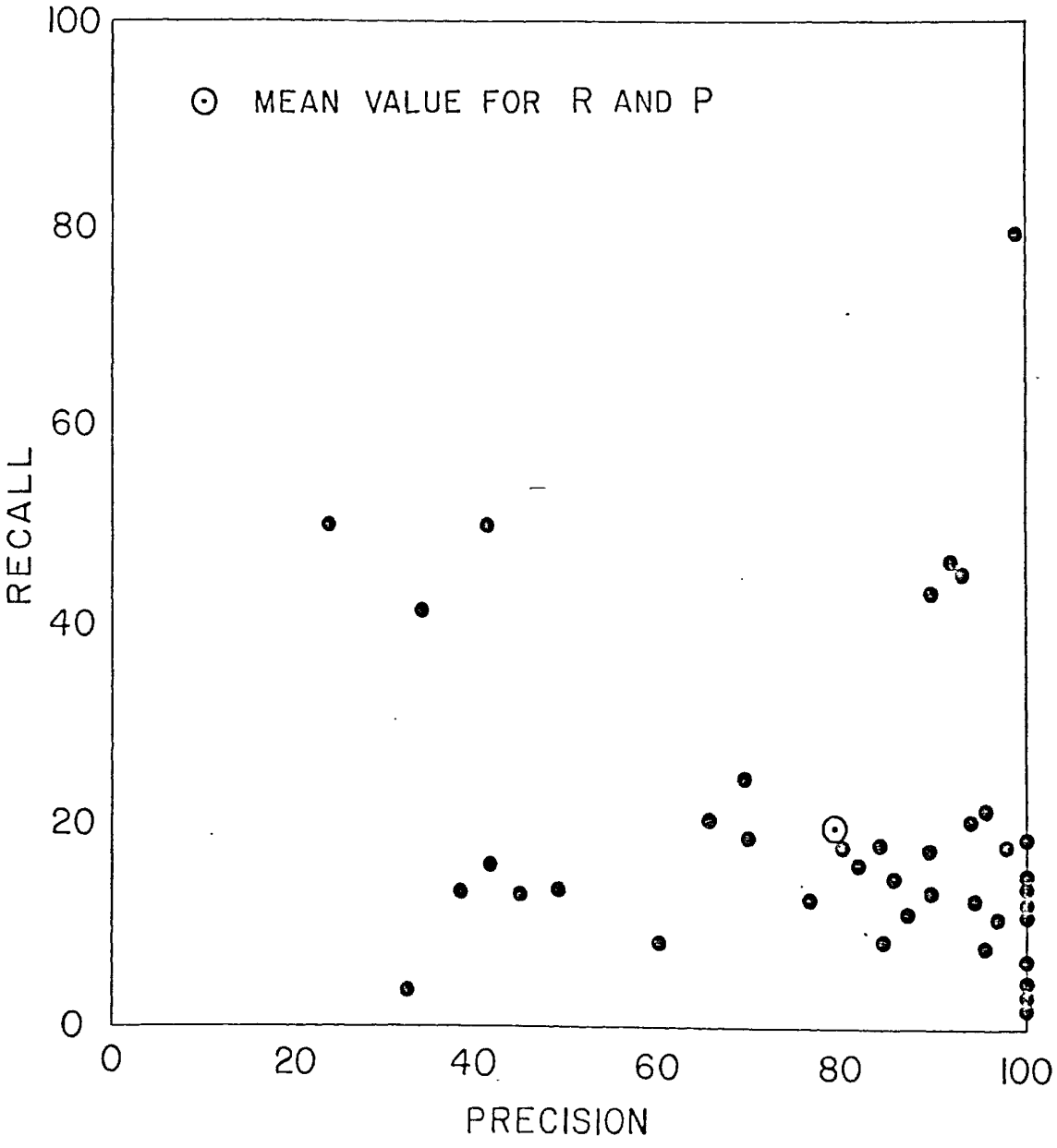

Figure 64

Another area of interest concerned the revisions made to information 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% of the total) were significantly different from the overall mean values, we might be able to infer something about the requesting procedure. Unfortunately, the values for Recall and Precision (23.9% and 62.1% respectively) for the substantially revised queries did not indicate a statistically significant difference from the mean overall values for Recall and Precision.

Finally, we tested the hypothesis that extremely high values of Precision for the retrieved sets would correlate directly with the lawyer's judgment of satisfaction with that set of documents (this 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 with the mean Precision for all requests. Although the Precision of the requests which were not revised significantly came out to be 85.4%, again the results were not statistically significant at the .05 level.

Retrieval Effectiveness: Lawyers vs. Paralegals

Consider the following argument: 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, the retrieval effectiveness can be significantly improved if the person who originated the information request was himself doing the searching at the terminal. This means that if a lawyer worked directly on the query formulation and query modification at the STAIRS terminal, rather than use

his paralegal as an intermediary, the retrieval effectiveness would be improved.

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

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

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

Test Results: Discussion

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

Why Was Recall so Low?

The low values of Recall occurred because full-text retrieval is very difficult to use to retrieve documents by subject. Full-text retrieval is very difficult to use because its design is based on the assumption that it is a simple matter for inquirers to predict the exact words and phrases used only in the text of documents that they would find useful. This assumption is not new, it goes back over 25 years to the early days of computing. The basic idea was that one could use the formal aspects of text to predict its meaning or subject content. The formal aspects of text are the physical and observable aspects of text such as the occurrence, location, and frequency of words, and

(to the extent that it could be precisely described) the syntactic structure of word phrases. The hope that motivated work on this problem was that by exploiting the high speed of a computer to analyze the formal aspects of text, one could get the machine to deal with text in a "comprehending-like" way; i.e., to identify the subject content of texts. This field of endeavor is called "Automatic Indexing" or, in a more general sense, "Natural Language Processing." During the past two decades many computer 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 [6]. The results of these experiments have shown that full-text document retrieval has only worked well on unrealistically small databases. But the belief in the predictability of the words and phrases that might be used to discuss a particular subject is a difficult prejudice to overcome. It is not a stupid prejudice. In a naive sort of way it is very appealing. But it is a prejudice nonetheless, because up until this study the effectiveness of full text retrieval has not been substantiated by reliable Recall measures on realistically large databases. The difficulty with full text retrieval systems can be stated quite succinctly: It is impossibly difficult for inquirers to predict the exact words, word combinations, and phrases which are 1. used by all (or most) relevant documents, and 2. used only (or primarily) by those documents. (Observation 2 is most frequently overlooked by proponents of full-text retrieval.) The following brief (but typical) examples will illustrate this point. One issue with which the lawyers were concerned was an accident which had occurred and was now an object of litigation. The lawyers wanted all the reports, correspondence, memoranda, and minutes of meetings which discussed this accident. Formal queries were constructed which

contained the word "accident(s)" along with several relevant proper nouns. Later in our search for unretrieved relevant documents we 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 proper names involved (because they were obvious to those discussing the issue). The manner in which an individual referred to the accident 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--they called it an "accident." But those individuals who were personally involved in the event (and, perhaps, culpable) tended to refer to it euphemistically. It was they who referred to the accident as, inter alia, an "unfortunate situation," or a "difficulty." But these were not all the terms which were used on relevant unretrieved documents. Sometimes the accident was referred to obliquely as "the subject of your last letter," or "what happened last week was...," or, as the opening lines of the minutes of a meeting discussing the issue began "Mr. A: We all know why we're here...". Sometimes relevant documents dealt with the problem by only actually mentioning the technical aspects of why the accident occurred, and not mentioning the accident itself or the proper names involved. In addition, much relevant information discussed the situation prior to the accident, and, naturally, contained no reference to the accident itself.

Another information request identified three key terms or phrases that were used to retrieve relevant information, but later we were able to find 26 other words and phrases which retrieved additional relevant documents. The three original key terms could not have been used individually because they would have retrieved 420 documents, or approximately 4,000 pages of hard copy, an unreasonably large retrieved set most of which contained irrelevant

information. Another information request identified four key terms/phrases that were used to retrieve relevant documents, but later we were able to use 44 additional terms and combinations of terms to retrieve relevant documents that had been originally missed.

Sometimes we could follow a trail of linguistic creativity through the data base. In one example, one of the key phrases was "trap correction." This, of course, was used to retrieve relevant documents, but later we discovered that relevant, unretrieved documents had discussed the same issue but referred to it as the "wire warp." We continued our search and found that in other documents this same thing was referred to in a third way: The "shunt correction system." Further, we discovered that the inventor of this system was a man named "Coxwell." This directed us to some documents he had authored discussing this system, only he referred to it as the "Roman circle method." Using this phrase as a formal query we discovered still more relevant unretrieved documents. But this wasn't the end. Further searching revealed that this 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 our search ended (having taken 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.

Since the database included many items of personal correspondence and 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 to effective retrieval. Key search terms

(which were essential parts of phrases) such as "flattening," "gauge," "memos," and "correspondence," were used in formal queries to retrieve relevant documents. But we were also able to retrieve relevant documents using the same phrases but with the search terms spelled "flatening," "guage," "gage," "memoes," and "correspondance," respectively. Such misspellings are tolerable in normal everyday correspondence, but when included in a computerized database they become literal traps for inquirers who must not only anticipate the key words and phrases which might be used to discuss an issue, but also all the possible misspellings, letter transpositions, and typographical errors which might be made in using those key words and phrases (and we make no claim to having anticipated all the possible errors).

Some of the information requests placed almost impossible demands on the ingenuity of the individual who constructed the formal query. In one situation, the lawyer wanted "Company A's comments concerning...". Just looking at the documents authored by Company A was not enough. Many relevant comments were not retrieved initially because these comments were embedded in the minutes of meetings or recorded second-hand in the documents authored by others. Merely 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). But predicting the exact phraseology of the text in which Company A commented on the issue was almost impossible. Examples which occurred in unretrieved relevant documents included "Co. A agreed to consider," "Co. A. said," or "Co. A pointed out that." Sometimes Company A was not even mentioned, it was merely noted that So-and-so (who represented Company A) "said/considered/remarked/pointed out/commented/noted/explained/discussed," etc.

In some information requests the most important terms and phrases were not used at all on 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 found which did not report steel quantity at all, but merely recorded the number of such things as "girders," "beams," "frames," "bracings," etc. In another request it was important to find documents which discussed "non-expendable components." Here, 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 "non-expendable."

These examples are only a few of the myriad linguistic problems which confronted the inquirers who had to use STAIRS to search for relevant textual information. The task was an impossibly difficult one, due to the unlimited and unpredictable ways in which individuals can talk about a particular subject.

We should pause here to answer the second question which was prompted by our evaluation. Namely, why didn't the lawyers realize that they weren't 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 data base, but 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 weren't 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 4 individuals (2 lawyers and 2 paralegals) to have total recall of all the

important supporting facts, testimony, and related data which were germane to the 50 issues for which they submitted information requests. If they had such recall they would have no need for a computerized, interactive retrieval system. It is a well known fact 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.) 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. But because of the linguistic issues which we touched on earlier, there was certainly reason enough for educated information system designers to suspect that the theoretical assumptions on which STAIRS (or any full-text retrieval system) was designed were questionable, and this is what prompted our study. The lawyers, in using STAIRS, were in a position of trust. What they should have been provided with was a retrieval system whose theoretical foundations were solid enough to assure them that they were getting the best Recall to be expected. Full-text retrieval, for theoretical reasons, simply does not provide this.

Deterioration of Recall as a Function of File Size

One of the reasons 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 (query formulation and revision, and retrieved set evaluation) required to obtain the same Recall increases as the database increases, and this increase in searching effort may increase at a faster rate than the increase in the database size. Let's consider a simple example. On the database we studied there were many search terms which, used by themselves, would retrieve over 10,000 documents. We shall call the retrieval of intolerably large sets of documents "output overload." This is a frequent problem of full-text retrieval systems.

A retrieved set of several thousand documents would be impractical to browse for relevant information, so the inquirer would be forced to reduce this "output overload" by reformulating the single term query so that it will retrieve fewer documents. He can do this in the following way: If a single term query w_1 retrieves too many documents, then he could 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," to include all the STAIRS variations on a Boolean conjunction). The reformulated query cannot retrieve more documents than the original query; most probably, it will retrieve many fewer documents. Thus, the process of adding intersecting terms to a query can be continued until the size of the output has been reduced to some manageable number. (This strategy, and its consequences are discussed in more detail in [7].) But what does this do to Recall? When an inquirer attempts to narrow the size of the output by adding intersecting terms the value of Recall goes down because, with the addition of each term, there is a probability that some relevant documents will be <u>excluded</u> by that reformulated query.

Let's look at the deterioration of Recall from a probabilistic point of view. For each query, there is a class of relevant documents which we can

$P(Sw_1)$ = Probability Searcher uses term w_1 in a search query

$P(Sw_2)$ = Probability Searcher uses term w_1 in a search query

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

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

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

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

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

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

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

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

e.g.: $P(.6) \times P(.7) \times P(.5) \times P(.6) = .126$

Table 3

designate as R. We can 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. As a result, 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.

But 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 query. The probability that he 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. But these probabilities must be multiplied by the probability that the searcher will select w_2 as part of his query, and the probability that w_2 will occur in a relevant document. Thus, calculating the 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 3). When we consider a three or four term query the value of Recall drops off even more sharply. The problem of output overload is an especially critical one in

full-text retrieval systems such as STAIRS. The frequency of occurrence of search terms in the STAIRS database is considerably larger than (and increases at a greater rate than) the frequency of occurrence (or "breadth") of index terms in a database where the terms are manually assigned to documents. Hence the user of a full-text retrieval system will be more quickly confronted with the problem of output overload than the user of a manually indexed system. The solution that STAIRS offers is for the inquirer to reformulate his query using such conjunctive connectives as "and," "adjacent," "with," and "same." This strategy will, in fact, reduce the number of documents retrieved to a manageable number. However, it will also eliminate relevant documents with each successive intersection of a new term, and the value of Recall will go down. Search queries employing 4 or 5 intersecting terms were not uncommon among the queries we looked at. But the probability that a query which intersects 5 terms will retrieve relevant documents is quite small. For example, if we were to assign a probability of .7 to all the respective probabilities in a hypothetical 5 term query (and .7 is a very optimistic average value) such as we did in the 2 term query in Table 3, we would find out Recall level for that query would be .028. In other words, that query could be expected to retrieve less than 3% of the relevant documents in the database. If the probabilities for the 5 term query were a more realistic average of .5, the Recall value for that query would be .0009! That is, if there were 1,000 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 wants to retrieve a high percentage of the relevant documents.

Discussion

The reader may be surprised that the results of this test of retrieval effectiveness are so striking, and he is not alone. The lawyers who participated in the test were equally astonished by the outcome. Although there are sound theoretical reasons why we should expect these results, such results seem to run counter to previous tests of retrieval effectiveness for full-text retrieval. Is there a discrepancy here? Let's look more closely at some of these earlier studies.

Two pioneering evaluations of full-text retrieval systems come to mind, representing high water marks in such endeavors. Swanson [1] and Salton [2] are respected researchers in the field of information retrieval in general and document retrieval in particular. Their respective evaluations of full-text retrieval systems 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, has shown that full-text document retrieval does not operate at satisfactory levels, and that there are sound theoretical reasons to expect that manually indexed document retrieval systems will be more effective than full-text systems. Who's 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 analysis done on small experimental databases of less than 750 documents. Our study was done not on an experimental database but on an actual operational database of almost 40,000 documents. Had Swanson and Salton been fortunate enough to be able to study a retrieval system as large as ours they would have undoubtedly observed phenomena similar to what we discovered (Swanson [8] was later to comment perceptively on the difficulty of drawing accurate

conclusions about document retrieval from experiments using small databases). It has only recently been observed that information retrieval systems do not scale up [9]. That is, retrieval strategies which work well on small systems do not necessarily work well on larger systems, primarily because of the problem of output overload--a problem which does not exist on small systems. 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. But large-scale, detailed retrieval effectiveness studies, like the one reported here, are unprecedented because they are incredibly expensive and time-consuming (the experiment took 6 months, involved two researchers and 6 support personnel, and, when considering all direct and indirect expenses, cost over half a million dollars to conduct). Nevertheless, Swanson and Salton's earlier full-text evaluations remain pioneering studies of retrieval effectiveness and, rather than contradict our findings, have an illuminating value of their own.

At this point it will be useful to consider an objection which might be made to our evaluation of STAIRS. This objection claims that the low Recall observed in the evaluation was not caused by STAIRS, but was due to query formulation error by the individuals who used the system. This objection arises from the realization that, in principle, virtually any subset of the database is retrievable by some simple or complex combination of search terms. The task of the searcher is merely to find the right combination of search terms to retrieve all and only the relevant documents. But the searchers should not bear the blame here. Perhaps an analogy may be illuminating. 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, for you to find the correct combination. But it may be impossibly difficult to do so in practice. A full text retrieval system must bear the burden of retrieval failure because it places the inquirer in the position of having to find (in a relatively short time) an impossibly difficult combination of search terms to retrieve relevant documents. Such search queries are difficult to formulate because the natural language text which must be searched is unpredictably varied and creative in the way that it may discuss the same subject. The inquirer who must use a full text retrieval system to find relevant information on a realistically large database is in the same unfortunate position as the individual who must find the combination to a lock. True, we, as evaluators, did find the "combinations" of search terms necessary to retrieve many of the unretrieved relevant documents, but three things should be kept in mind: 1. We make no claim for having found all the relevant unretrieved documents. We may not have even found half of them (our sampling technique covered only a small percentage of the database). 2. A tremendous amount of search time (sometimes over 40 hours of on-line time) was involved with each request (the entire test took almost 5 months). Such inefficiency is clearly not in consonance with the high speed desired for computerized retrieval systems. 3. The evaluators combined over 40 years of practical and theoretical experience in information system analysis, and could be expected to have somewhat better searching abilities than the typical STAIRS searcher. The evaluators could find many unretrieved relevant documents, but STAIRS is sold under the premise that it is easy to use and requires no sophisticated training on the part of a potential searcher. Yet this study is a clear demonstration of how sophisticated the inquirer's search skills must be to use STAIRS (or, mutatis mutandis, any other full text retrieval system) at all. In

short, the full-text retrieval system user is in the same position as an individual who buys a new car only to find out that it takes a driver as skilled as Jackie Stewart to drive it. (There is some evidence that recognition of this problem is beginning among at least one full-text retrieval vendor, WESTLAW, which has made its reputation by offering full-text access to legal cases. Because of recognized difficulties with full-text retrieval, WESTLAW has now begun to supplement its full-text retrieval with manually assigned index terms for its cases.)

Summary

This paper has described and discussed a major, detailed evaluation of a full-text document retrieval system. We have shown that not only did the system not work well in the environment in which it was tested, but that there are theoretical reasons why full-text retrieval systems are unlikely to perform well in any retrieval environment. The optimism of early studies which examined full-text retrieval was biased by the small size of the databases used in the experiments. The superiority of document retrieval based on assigned index terms over full-text retrieval is not apparent until databases are of a more realistic, real-world size. Previous studies have only been concerned with showing that full-text search is competitive with searching based on manually-assigned index terms. If it were competitive, then, it is argued, full-text retrieval would save one from the cost of indexing. But does a full-text retrieval system really save the purchaser time and money? Can you really get something for nothing? It is true that you don't have to do any manual indexing or vocabulary construction in a full-text retrieval environment, but are there any costs which a full-text system incurs which a manual system does not? Yes there are. First, there is a substantial

increase in time and cost to input 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 must incur 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 in time alone would more than compensate for the "additional" time needed for manual indexing and vocabulary construction. But this isn't the only savings which a manually indexed system enjoys. The 20 fold increase in document record size would mean the database for a fulltext system would be perhaps 20 times larger than a manually indexed database. This increase in database size would result in increased storage and searching costs for a full text system. Finally, because the average number of searchable subject terms per document for the full-text retrieval system described here was approximately 500, while a manually indexed system might have a subject indexing depth of about 10, the dictionary which must list and keep track of all these assignments (provide pointers to the database) could be as much as 50 times larger on a full-text retrieval system as on a manually indexed system. We can see now that a full-text retrieval system does not give us something for nothing. There are significant hidden costs for such a design selection. 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."

Acknowledgements

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

REFERENCES

1. D. G. Swanson, Science 132, 1099 (1960).

2. G. Salton, Science 168, 335 (1970).

3. P. Zunde and M. E. Dexter, Amer. Doc. 20, 259 (1969).

4. T. Saracevic, J. ASIS [Amer. Soc. Inf. Sci.] 26, 321 (1975).

5. J. A. Swets, Science 141, 245 (1963).

6. K. Sparck Jones, Automatic Keyword Classification for Information Retrieval (Butterworths, London, 1971).

7. D. C. Blair, J. ASIS [Amer. Soc. Inf. Sci.] 31, 271 (1980).

8. D. R. Swanson, Library Quarterly 47, 128 (1978).

9. H. L. Resnikoff, Bull. ASIS [Amer. Soc. Inf. Sci.] 5, 27 (1978).