UNITED STATES
NUCLEAR REGULATORY COMMISSION
WASHINGTON, D.C. 20555

INFORMATION PAPER ON ABSTRACTING IN THE LICENSING SUPPORT SYSTEM

Office of the Licensing Support System Administrator

September 12, 1990

September 12, 1990

LSSA INFORMATION PAPER ON ABSTRACTING IN THE LICENSING SUPPORT SYSTEM

I. PURPOSE OF THIS PAPER:

At the upcoming October, 1990 meeting of the NRC Licensing Support System Advisory Review Panel (LSSARP), the members are scheduled to continue the discussion on their recommendation to the LSS Administrator (LSSA) on the content of the LSS Header. One open item was the extent to which documents in the LSS should be abstracted. The purpose of this paper is to lay out information about abstracting which the LSSA believes should be taken into consideration by the LSSARP members as they examine this issue.

II. BACKGROUND:

During the March 1990 meeting of the LSSARP, a Technical Working Group was formed to prepare a draft recommendation for the fields for the LSS Bibliographic Header and Full Header. The Working Group met several times and prepared a report to the full LSSARP. The report recommended that abstracts be required only for documents and non-documents that will not be available in searchable full-text (i.e., those with either header only or header and image only). The report further recommended that the abstract field be optional for documents that will be available in searchable full-text. The Technical Working Group determined that the LSSARP should discuss the issue as to which LSS document types or groupings should be abstracted.

During the June 7, 1990 meeting, the LSSARP members agreed that abstracts were required for materials that will not be available in searchable full-text. They then discussed at length the need for an abstract for LSS documents that will be stored in searchable full-text. These discussions centered around cost versus benefit considerations. Differing points were made about:

- the need for any abstract in the header, given availability of full text,

- the sizable cost of abstracting, and

- whether only selected sets of documents might need to be abstracted and, if so, which sets.

No firm recommendation evolved. To focus the issue and to provide more definitive information about the cost implications of alternative abstracting scenarios, the LSSA offered to prepare an issue paper for the members to consider prior to the next LSSARP meeting in October. Since the June LSSARP meeting, the LSSA staff has reviewed existing information science studies related to this issue and gathered industry data on the costs of abstracting. The following is the result of that investigation, including a discussion of abstracting options and some alternatives to abstracting.

III. ABSTRACTING -- WHAT IS IT?

A. TYPES OF ABSTRACTING

In the Library/Information Science discipline, three types of abstracts have evolved. All are based on the human review and summarization of the content of a document. In order of increasing depth and coverage, they are:

- ANNOTATIVE -- A short description of the document which briefly describes the subject, usually limited to a few lines in length. This type of abstracting can be done by the same staff doing the bibliographic or descriptive cataloging.

- INDICATIVE -- A longer description than the annotative abstract, giving a more detailed summary of the document scope and content. These abstracts are traditionally about 200 words in length. This type of abstracting is usually done by professional indexers/abstracters having subject matter background and/or experience. The documents are usually reviewed once both for the assignment of subject terms and for the development of the abstract.

- INFORMATIVE -- The most substantive type of abstracting which includes not only indicative information but also summarizes the findings, answers, or data in the document. Such abstracts often eliminate the need to obtain or read the entire document. The length varies based on depth of document content. As with the indicative abstract, this type of abstracting is also done by professional indexers/abstracters having subject matter background and/or experience.

However, unlike the Indicative Abstracts, this type of abstracting may or may not be done by the same staff that are subject indexing the documents. If not, then another staff resource is required.

It is obviously more expensive as one moves from annotative to informative abstracting because of the additional time and higher level of expertise involved in reviewing the document and composing the abstract. Section IV and Appendix A. contain more information on the cost of abstracting.

B. ABSTRACTING IN THE LSS ENVIRONMENT

Given that the LSS Title/Description field is intended to contain (a) the titles of formal publications or (b) a brief description of less formal or untitled documents, all LSS documents will have **annotative**-type abstracts. This makes the assumption that titles of publications are somewhat descriptive of content. Therefore, annotative abstracting is not considered from a benefit-costs perspective in this issue paper.

Also, in the opinion of the LSSA, the LSS should not attempt under any scenario to provide **informative** abstracts because (1) the costs are excessively high and (2) such treatment of LSS documents is unwarranted given the availability of the document text on-line. The LSS abstract would only be intended as a search aid, not as a surrogate for the document itself, which is often the case with systems providing informative abstracts.

Therefore, in discussing the pros and cons of abstracts in the LSS environment, this paper assumes that any abstracts would be of the **indicative** type.

C. BENEFITS OF INDICATIVE ABSTRACTS

The following is a list of the potential or reputed benefits of having an abstract field in a full-text database. Where applicable, we have included a summary of the information gained from relevant research studies. It should be noted that no specifically applicable research has been found that directly speaks to the benefits/costs of abstracts in a full-text database having keyterms and header data, such as will be the case with the LSS.

1. **IMPROVED PRECISION** -- The presence and use of abstracts may improve the precision of subject/content searches because it is assumed that if a word or phrase is in the abstract, then it is probably a primary topic of the document. This

precision is gained by limiting word/phrase searches to the abstract field, either initially or after retrieving a document set via search of full-text or other parameters.

There is a current on-going debate in the information science literature about the benefits and power of full-text database software as compared to traditional systems that have only bibliographic (fielded) data, subject indexing, and abstracting. Most of this debate centers around the balance of "recall" versus "precision" capabilities. The attached articles are representative of the discussions and data surrounding this debate (see Attachments #1 through #5).

It is known that in striving to achieve the greatest **recall** (retrieval of all relevant documents), the **precision** (retrieval of only relevant documents) of search results suffers. This axiom is applicable to all types of information systems, ranging from bibliographic only to full-text systems. However, the degradation of precision to assure greatest recall is magnified in large full-text systems, especially for collections on a narrow and/or homogeneous topic, such as the HLW LSS. This problem will be further exacerbated in the LSS environment of decision support and litigation support where knowledge of all relevant materials appears more to be essential.

In a 1986 article (Attachment #1), Gerald Salton summarizes the results of several related studies. Simplistically presented, the precision/recall performance of different access methods can be drawn from two of the studies. These data support the belief that searching the abstracts can significantly improve recall (as compared to searching the full-text alone without) a significant loss in precision.

Searching the:	Recall Ratios*	Precision Ratios*
a. Text of Abstract	0.78	0.63
b. Controlled Descriptors Subject Indexing	0.56	0.74
c. Full Document Text	0.20	0.75

* **Recall Ratio** is number of retrieved relevant documents as percentage of all of the relevant documents in the database.

Precision Ratio is the number of retrieved relevant documents as percentage of all retrieved documents

As indicated in line b. above, the recall ratios are better if one has controlled subject terms to search as well as the full-text, without any significant loss of precision. Subject indexing will be done in the LSS.

2. **RELEVANCY REVIEW** -- Abstracts provide a summary of the entire document. Therefore, browsing the abstracts of a retrieved set of documents can aid in determining the usefulness of the document and the context in which the subject is treated without having to roam around in the text.

   Also, abstracts can be very helpful when reviewing document listings or bibliographies in hardcopy away from the LSS workstation. This would be the case when LSS search specialists or intermediaries, e.g. librarians, research assistants, and paralegals, are performing searches in response to "client" requests. In one study, the presence of an abstract reduced the number of "missed documents" -- documents judged as not relevant by a review of the titles only, but which were subsequently determined as relevant after a review of the abstracts (Attachment #6).

3. **COST SAVINGS** -- Abstracts can potentially reduce the need for printing hardcopy of documents if a review of the abstract is sufficient for the searcher to determine the relevancy of the document for his/her needs.

4. **TIME SAVINGS** -- Abstracts can reduce on-line time if, as above, review of the abstracts negates the need to browse/read the full-text.

D. LIMITATIONS:

1. Abstracts are only as good as the abstracter. They are subjective, whether it be the author's characterization of his/her work or the abstracter's interpretation of the author's work.

2. Abstracts do not improve recall of subject/content searches in a full-text database if the abstract does not contain different terminology from the text. Different terminology that could improve recall might be more generic, more specific, synonyms, or the translation of jargon.

3. Abstracting only certain document types/categories places a burden on the user to know when abstracting was done and when it was not. Otherwise, users could unknowingly formulate search strategies that would provide false results. For

example, if all documents in a collection are not abstracted, then searches limited to the abstract field will automatically exclude non-abstracted documents and thereby possibly exclude relevant materials from the resulting hitlist.

IV. COSTS OF ABSTRACTING

A. AVERAGE COST PER ABSTRACT

The LSSA collected abstracting cost and productivity information from six companies that perform abstracting services. The information provided by respondents varied in terms of assumptions, such as variations in the size of documents, the QC reviewers/supervision ratios, and scope of abstracting. It was therefore difficult to normalize the data. However, there was not such a disparity in the data that some useful figures could not be compiled. The assumptions used for this paper are listed in the Table below and Appendix A.

Data was also provided by SAIC, based on their experience in the LSS prototype cataloging efforts. Their data show abstracting times of about seven (7) minutes per document based on a sample of 47 documents, each averaging 48 pages. Unfortunately, the SAIC timing estimates did not include a quality control review. Also, it was uncertain whether these times consistently included the actual review and analysis of the document scope and content before the composition and keying of the abstract.

B. ESTIMATED COSTS IN THE LSS

The following table presents the estimated costs of abstracting LSS documents by document type. The figures on the number of documents are extrapolations from recent SAIC re-evaluations of the size of the LSS database (see Attachment #7). The estimated number of pages in this SAIC report was divided by nine (9) to develop an estimated number of documents. The figure of nine (9) pages per document was selected because this was the size of the average document in the DOE Nevada RIS collection, which will contribute the vast majority of documents to the LSS.

The distribution of the estimated number of documents by major document types is based on recent figures from the three major HLW document collection: DOE's RIS systems in Las Vegas and at DOE Headquarters and the NRC's NUDOCS system.

Even though the figures in the table below are just gross estimates and may differ from the actual volume/costs experienced in the future; these figures are based on the best available data. For the purposes of this paper, they do provide the LSSARP members with a significantly improved basis for decision making.

Table 1. ESTIMATED COSTS OF ABSTRACTING IN THE LSS
(Numbers of Documents & Dollars in thousands)

LSS DOCUMENT COLLECTION BY DOCUMENT TYPE	Cumulative Document Counts and Costs by Specified Year					
	BY 1995		BY 2000		BY 2005	
	NO. OF DOCMNTS	EST. COSTS	NO. OF DOCMNTS	EST. COSTS	NO. OF DOCMNTS	EST. COSTS
TOTAL	1,278	$33,179	2,296	$59,595	3,759	$97,581
CORRESPONDENCE (64%) 3 doc/hour	818	$17,996	1,469	$32,318	2,406	$52,932
PUBLICATIONS/ REPORTS (23%) 2 doc/hour	294	$9,700	528	$17,427	864	$28,512
LEGAL & OTHER DOCUMENTS (13%) 2 doc/hour	166	$5,483	299	$9,850	489	$16,137

Assumptions:

1. A fully loaded rate of $66.00 per hour. This includes the costs of labor (abstracters, quality control reviewers, and supervisors), G&A, overhead, and fee. Abstracting work activities include reading documents, composing abstracts, keying in the abstracts, and performing quality control and supervision.

2. A production rate of two abstracts developed and reviewed per hour ($66.00 divided by 2 = $33/abstract) was used for the Publications/Reports and Legal/Other Document categories. This is the production figure used by the National Federation of Indexers and Abstracters for 200 word indicative abstracts. For correspondence with typically fewer pages than the other two categories, a production rate of three per hour was used ($66.00 divided by 3 = $22/abstract).

3. While it is acknowledged that a portion of the LSS documents, particularly formal publications, will have an abstract or summary within the body of the document, no cost reduction was factored into this table. This decision was based on responses of the surveyed abstracting companies. They were reluctant to reduce estimates even if documents contained abstracts, due to the time required to verify the quality of the existing abstract and to edit as required for consistency of coverage with other abstracts. This decision was also supported in the timing tests performed by SAIC in their prototype. Also, no adjustment was made to acknowledge that some documents, such as transmittal correspondence, would not warrant abstracting, given that an annotative summary would be contained in the Title/Description field.

V. ALTERNATIVES TO ABSTRACTING

Section III.C presented the potential benefits of having abstracts in the LSS. This section highlights some of the LSS features currently specified in the SAIC draft LSS Search and Image Design Document which will provide some of the same benefits of abstracting without the continuing costs of abstracting. These software features, if not part of the off-the-shelf database package, can be developed at a finite, one time cost. This section also discusses some other features that could increase precision and recall.

A. CURRENT DOE LSS DESIGN FEATURES

1. Header Field Analysis: After a searcher has developed a hitlist of documents based on his/her search statement, this optional feature, if invoked, would present to the user a computed table of the frequency of occurrences of values for any specified Controlled Vocabulary Header Field. This shows the distribution of Descriptors, Sponsoring Organizations, Author Organizations, etc. within their hitlist.

For example, given the best known search strategy, the user creates a hitlist of 230 documents on boreholes and volcanic rocks. The user then requests the Header Analysis feature, using the Descriptor field. The LSS system would then present a listing of all Descriptors used to describe the 230 and show the number of documents having each descriptor, in decreasing frequency order. The table would look something like:

This query found 230 units.
Header Analysis on Descriptor Field:

Descriptors	Frequency
Fractures	47
Fractures (Geologic)	43
Topopah Springs Member	39
Boreholes	36
Drill Cores	30
Stratigraphy	25
:	
:	
:	
Volcanic Rocks	11
Structural Geology	10
Strain (Geology)	4

The user could use this information about their hitlist to select parameters of greatest or least interest to refine the search statement and create a query with greater **precision**. For example, the searcher might now want to broaden the search to include all documents on Topopah Springs Member while also excluding documents on Stratigraphy and Strain.

2. Ranking Retrieved Documents Based on Selected Term Frequency: This LSS feature will allow the user to rank and display the documents in his/her hitlist in decreasing order according to density of selected ASCII-text words in the text. Density is defined as the number of times a relevant words or phrases appear in the document as a percentage of the total number of words in the document. For example, the words abstracts, abstracted, abstracting, and abstracters are repeated about 140 times in this 4,000 word paper. This represents 3.5% of all words in this paper. The percentage would be even greater if "stop" words (such as a, the, were, most, in, etc.) were excluded from the total word count. This process will present the hitlist in an order which provides the most **relevant** documents first on the assumption that if the specified words are repeated frequently in the document, that is a major topic covered in the document.

B. POTENTIAL LSS DESIGN FEATURES

The following are search and retrieval software features that are not currently in the DOE design. These features may warrant further investigation, given the costs of abstracting, the concern of excessively large hitlists, and the problems of low recall and low precision in large text databases.

1.a. Automatic Abstracting -- There are current software packages that purport to scan existing text and present the contents into an abstract-like summary. Such a software feature could be used to add a summary to the LSS header record for presentation to searchers and reviewers of bibliographies to enhance their determination of the relevance of documents retrieved. This would potentially provide the benefits of: (a) reducing the orders for non-relevant documents or (b) finding relevant documents that might have judged non-relevant upon review of the bibliographic information only.

1.b. Optional Extensive Bibliography Format -- LSS users could the have option of ordering the "first" ASCII page of each document in their hitlist to be printed along with a header bibliographic listing. Such a feature would have the same benefits as Automatic Abstracting, described above.

2. Sophisticated Ranking Algorithms -- Over the past several years, the information science literature has contained many articles about research to improve text search results using a variety of statistical and lexical analysis methods. Basically, these are centered on the clustering of related or synonymous terms

and word patterns. Attachments #4 and #8 are examples of such techniques. The capabilities of such software enhancements to improve recall and precision will be carefully monitored. As features become proven, they could be incorporated into the LSS design over the life of the system.

VI. PROS & CONS OF DIFFERENT OPTIONS FOR ABSTRACTING:

A. ALL DOCUMENTS

PROS: ▸ Consistency and simplicity

CONS: ▸ Prohibitively Expensive

▸ Not warranted for traditional 'correspondence' given:

- use of Title/Description Field which will provide short annotative summary for relevancy review.
- full-text search capability
- multiple other access points in the header fields for content/subject searches of all documents, such as descriptors, identifier, project/special class fields etc.

B. ALL NON-CORRESPONDENCE-TYPE DOCUMENTS - "everything but .."

Exclude letters, memos, telephone conversation reports...

B.1 Abstract all non-correspondence regardless of how long or short the document.

PROS: ▸ Less expensive than Option VI.A.

CONS: ▸ Somewhat wasteful given that some "short" documents do not warrant such treatment.

B.2 Abstract only non-correrpondence over a certain page count.

PROS: ▸ Less expensive than VI.B.1.

- Increased benefits of <u>relevancy review</u> and <u>precision</u>

CONS:
- Selection of document size cutoff is arbitrary and subject to debate.
- Searchers are very <u>unlikely</u> to keep this arbitrary rule in mind. Therefore, if they limit their searches to the Abstract Field for <u>precision</u>, then they could unknowingly exclude whole sets of documents and get erroneous search results.

C. ABSTRACT ONLY SPECIFIC DOCUMENT TYPES.

C.1 For All Documents Coded as Specified Document Types -- Pick up Abstracts/Summaries as available within documents or compose and add if not.

PROS:
- Less Subjective or arbitrary in the selected universe than VI.B.2.
- Much less expensive because of smaller universe of documents to be abstracted.
- Most understandable alternative to most, if not all, searchers. Therefore least likely to be misused in searching.

CONS:
- Still somewhat subjective in that the assignment of Document Type codes is somewhat subjective.
- Inconsistent treatment of abstracts and therefore varying quality if abstracts drawn from the text are not strictly reviewed for consistency with LSS abstracting standards.

C.2 Only Store Abstracts in Headers for Documents which have author-generated Abstracts/Summaries available in the text which can be "grabbed" and put in header as searchable full-text.

PROS:
- The least expensive alternative while still allowing searching of this text because submitter's preparation staff and/or LSSA staff do not have to compose and enter the abstract.

▸ The abstract listed in bibliographies will assist the reviewer in determining the potential relevance of documents retrieved.

CONS: ▸ **Universe of documents which contain abstracts for searching and for presentation is totally random. This does not appear to be a viable option because searchers could not use these randomly existing abstracts with any reliability for identifying relevant documents.**

▸ Subjective in determining if document contains text which could be used as an abstract.

▸ Inconsistent treatment of abstracts and therefore varying quality if abstracts drawn from the text are not strictly reviewed for consistency.

C.3 <u>Only Store Abstracts in Headers for Documents which have author-generated Abstracts/Summaries available in the text which can be "grabbed" and put in header but not allow this Abstract field to be searchable.</u>

PROS: ▸ The least expensive alternative. A minimal cost to transfer and store the pre-existing text in the header in a non-searchable field.

▸ The abstract listed in bibliographies will assist the reviewer in determining the potential relevance of documents retrieved.

▸ By not allowing searches to be limited to Abstract Field in this option, it prevents users from unknowingly eliminating potentially relevant sets of documents.

CONS: ▸ This option presents a design issue to be solved because the abstracts in LSS header records that describe documents or data that are <u>not</u> stored in searchable full-text would have to be made searchable.

VII. <u>CURRENT LSSA STAFF VIEW:</u>

The LSSA staff believes strongly that manually prepared abstracts should not be created for inclusion in the Licensing Support System in searchable text for those documents that are already stored in searchable full-text due to the substantial costs projected for abstracting in comparison to the benefits. Although there is the potential for low recall and precision ratios in large text databases, abstracting is not the only remedy. The other access points in the LSS header fields and the software features specified in the current LSS design will greatly enhance to searchers ability to create useful sets of documents. Also, the LSSA staff will continue to work with DOE in investigating additional software tools to increase performance and will recommend the development of such software if it is a cost-effective approach.

The LSSA staff does believe that the text of abstracts that already exist in documents should be captured in the Full LSS Header. This would be in a non-searchable field to be used for presentation and relevance review only, (Option C.3) above. This assumes the design issue can be solved related to the need to search abstracts for those documents/data not stored in searchable text.

APPENDIX A

SUMMARY OF INDUSTRY SURVEY OF ABSTRACTING COSTS

DIRECT HOURLY LABOR RATES	COMPANY A	COMPANY B	COMPANY C	COMPANY D	COMPANY E	COMPANY F	NFAIS
ABSTRACTERS	$13.50 - 18.00	$25.00	$10.00 - 15.00	Unit Charge	nr	$12.00	$13.50
QUALITY CONTROL REVIEWERS	nr	$25.00	nr	"	nr	nr	nr
SUPERVISORS	$30.00	$25.00	nr	"	nr	nr	nr
RATIO OF QC PERSONNEL TO ABSTRACTERS	1:2	1:5	nr	1:3	1:4	nr	1:4
RATIO OF SUPERVISORS TO ABSTRACTERS	1:20	1:15	nr	1:15	Same Person as QC	nr	nr
UNIT CHARGE PER ABSTRACT	nr	$58.50	nr	$33.29	$16.77	nr	nr
TIME TO PRODUCE AN INDICATIVE ABSTRACT	20 Pages of doc. per hour	135 mins/ document	nr	49 mins/ 35 page document	37 mins/ 12.5 page document	nr	30 mins/ document

NOTES: nr = not reported
NFAIS = National Federation of Abstracters and Indexers

APPENDIX A cont.

CALCULATIONS OF FULLY LOADED HOURLY RATE

Average Direct Hourly Rate:

Abstractors	=	$15.75
QC Personnel	=	20.00
Supervisors	=	27.00

Ratio of QC Personnel to Abstracters	=	1:3.5
Ratio of Supervisors to Abstracters	=	1:15

Abstractor's hourly rate	$15.75	
+ portion of QC rate	5.71	($20 hourly rate for QC personnel divided by 3.5)
	$21.46	
+ portion of Sup.rate	1.80	($27 hourly rate for Supervisors divided by 15)
	$23.26	
+ Overhead (120%)	27.91	
	$51.17	
+ G & A (20%)	10.23	
	$61.40	
+ Fee/profit (8%)	4.91	
	$66.31	=== Fully loaded hourly rate for abstracting services.

APPENDIX B

ATTACHMENTS

#1 Salton, Gerald. "Another Look at Automatic Text-Retrieval Systems." Communication of the ACM. 29(7). 648-656. July 1986.

#2 Blair, David C. and M.E. Maron. "An Evaluation of Retrieval Effectiveness for a Full-Text Document-Retrieval System." Communications of the ACM 28(3). 289-299. March 1985.

#3 Tenopir, Carol. "Contributions of Value Added Fields and Full-Text Searching in Full-Text Databses." Proceedings of the National On-Line Meeting - 1985. Medford NJ: Learned Information, Inc., 1985. pp. 463-470.

#4 Ro, Jung Soon. "An Evaluation of the Applicability of Ranking Algorithms to Improve the Effectiveness of Full-Text Retrieval. I. On the Effectiveness of Full-Text Retrieval." Journal of the American Society for Information Science. 39 (2), 73-78. 1988.

#5 A. Jordan, John S. Letter to the Editor, Journal of the American Society for Information Science (JASIS) 40(3), 362-363. 1989

B. Lancaster, F.W. Letter to the Editor, JASIS 40(3), 362. 1989.

#6 Saracevic, Tefko. "Comparative Effects of Titles, Abstracts, and Full Texts on Relevance Judgements." Proceedings of the American Society for Information Science. Vol. 6 Oct.1-4, 1969. pp. 293-299.

#7 Science Applications International Corporation. Licensing Support System, Revised Data Scope Analysis. Draft. dated August 28, 1990.

#8 Deerwater, Scott et. al. "Indexing by Latent Semantic Analysis" Journal of the American Society for Information Science 41(6): 391-407. 1990.

ATTACHMENT #1 ATTACHED

Edgar H. Sibley
Panel Editor

Evidence from available studies comparing manual and automatic text-retrieval systems does not support the conclusion that intellectual content analysis produces better results than comparable automatic systems.

ANOTHER LOOK AT AUTOMATIC TEXT-RETRIEVAL SYSTEMS

GERARD SALTON

An automatic text-retrieval system is designed to search a file of natural-language documents and retrieve certain stored items in response to queries submitted by a user. Typically, each stored item is described by using—for content identification—certain words contained in the document texts, sometimes supplemented by additional related information. Queries are often formulated by using as search terms words from the text that are interrelated by the Boolean operators *and*, *or*, and *not*. The retrieval system is then designed to retrieve all stored texts identified by an appropriate combination of query words. A user interested in information about the design of small computers might formulate the query [(minicomputers *or* microcomputers *or* hand-held calculators) *and* (design *or* construction *or* architecture)]. The retrieval system would then extract, from the file, items containing the identifiers "design" and "minicomputers," or "construction" and "microcomputers." [8, 16]

The effectiveness of a retrieval system is usually evaluated in terms of a pair of measures, known as *recall* and *precision*. Recall is the proportion of relevant material actually retrieved from the file, while precision is the proportion of the retrieved material that is found to be relevant to the user's needs. In principle, a search should achieve high recall by retrieving almost everything that is relevant, while at the same time maintaining high precision by rejecting a large proportion of extraneous items. When this happens, both recall and precision values of the search are close to 1 (or 100 percent). In practice, it is known that recall and precision tend to vary inversely, and that it is difficult to retrieve everything that is wanted while also rejecting everything that is unwanted.

In particular, when very specific query formulations are used, few nonrelevant items tend to be obtained, but also relatively few relevant ones. That is, a very specific query formulation produces high-precision and hence, low-recall, performance. As the query formulation is broadened, more relevant items are retrieved, thus improving the recall, but also more nonrelevant ones, thereby depressing the precision. In the latter case, one obtains high recall, but also low precision. A compromise often reached in practice is using a query formulation that is neither too narrow nor too broad. However, when a choice must be made between recall and precision, most users choose precision-oriented searches where only relatively few items are retrieved, and the user is spared the effort of examining a large amount of possibly irrelevant material—the penalty attached to a high-recall search.

In automatic retrieval systems, both query formulations and document representations can be altered to reach the desired recall and precision levels through the use of recall-enhancing devices (e.g., term truncation) to broaden the document and query identifiers, and precision-enhancing devices

This study was supported in part by the National Science Foundation under grant IST 83-16166.

(e.g., term weighting) to make item identifications more specific. A list of typical recall- and precision-enhancing devices appears in Table I.

Term truncation consists of using truncated terms, or word stems, instead of the original complete terms, for query or document identification. A form like "analy" would encompass the notions "analyst," "analysis," "analyzer," etc.—having a broader scope than any of the complete words. Other recall-enhancing devices involve using terms that are synonymous or related to the original ones or broader and more general. Such terms are generally available in thesauri and term hierarchies or are suggested by users during the search operations.

Term weights enhance the search precision by distinguishing the better, or more important, terms from the less important ones. Such a discrimination may also help rank the output in decreasing order of presumed importance. Other precision-oriented devices involve using term phrases instead of single terms—for example, "computer programmer" instead of "computer"—and supplying narrower or more specific terms. Useful term phrases might be available in a dictionary, or could be formed from sets of single terms that cooccur regularly in a collection of documents.

Most automatic text-retrieval systems provide for the use of truncated terms and the addition of broader, narrower, and related terms. Automatically generated term weights may also be used to distinguish items containing the more highly weighted terms from those containing terms of lower weight.

A recent article by Blair and Maron examines the well-known automatic text-retrieval system STAIRS as applied to a collection of 40,000 full-text documents—equivalent to some 350,000 pages of text—to answer 40 different user queries [1]. In STAIRS, words are normally extracted from document texts for content identification. After text words have been broadened using truncation, each word may be supplemented by lists of synonyms supplied by the user. When synonyms are specified, a search based on a particular term automatically extends to the whole synonym list. The STAIRS system also includes a ranking feature that retrieves documents in decreasing order based on total document weights, which are calculated by adding the weights of the query terms contained in each retrieved document [6].

Although some features of the STAIRS system are not as attractive as they might be (e.g., a more reasonable term weighting system might produce better retrieval performance), STAIRS is certainly a state-of-the-art full-text-retrieval system, and its operations are typical of what is obtainable with existing operational automatic text search systems. In the STAIRS retrieval test conducted by Blair and Maron, an average precision value of about 75 percent (0.75) was obtained, and an average recall value of 20 percent (0.20). That is, for each of the 40 test searches, three out of four retrieved documents were in fact pertinent to the user queries, and approximately one-fifth of the total number of relevant items present in the collection were retrieved.

[illegible]

In this article, we will argue that not only is this level of performance typical of what is achievable in existing operational retrieval environments, but that it actually represents a *high order* of retrieval effectiveness. We will present some major experiments comparing automatic retrieval with manual, controlled vocabulary systems on large document collections. We then address the theories underlying automatic indexing and propose a basic blueprint for implementing effective automatic retrieval systems, emphasizing that the future lies in automatic and not in manual systems.

TABLE I. Typical Recall- and Precision-Enhancing Devices

Recall-enhancing devices (term broadening)	Precision-enhancing devices (term narrowing)
Term truncation (suffix removal)	Term weighting
Addition of synonyms	Addition of term phrases
Addition of related terms	Use of term cooccurrences in documents or sentences
Addition of broader terms (using term hierarchy)	Addition of narrower terms (using term hierarchy)

THE BLAIR AND MARON RETRIEVAL TEST

In the Blair and Maron test of the STAIRS system, searchers were able to extract from a large collection of 40,000 documents a substantial number of useful items: since only one of four retrieved items proved extraneous, the time consumed considering useless items must have been comparatively small. However, the searchers in the Blair and Maron test were lawyers and the materials being searched were legal documents, and because the Anglo-American legal system is based on the concepts of common law and judicial precedence, many lawyers are of necessity high-recall users. In this tradition, knowing how a particular legal case must be approached often means examining all possible previous cases that may be similar in some respect to the current case. The high-precision output obtained by Blair and Maron, which rejected most nonrelevant materials, but also obtained only about 20 percent of the potentially useful items, might be entirely suitable in another environment (e.g., for research workers, university professors, and students). However, in the case of the legal personnel that actually conducted the searches in the Blair and Maron test, a better recall performance was considered essential even at the cost of decreased search precision.

From their retrieval test, Blair and Maron derive three main conclusions [1]: First, they assert that, when high recall is essential in searching large collections, users cannot simply broaden the search request (as would be done experimentally for small collections) because of the problem of output overload. More specifically, they claim that, when broader search formulations are used, search precision may suffer intolerably, and users might be swamped with masses of irrelevant material. For this reason, the authors conclude that earlier test results showing the superiority of text-based retrieval over manual systems are not necessarily relevant to large, real-world collections.

Second, Blair and Maron argue that, when high recall is desired, manual indexing is preferable to full-text searching.

> . . . the full text system means the additional cost of inputting and verifying 20 times the amount of information that a manually indexed system would deal with. This difference alone would more than compensate for the added time needed for manual indexing and vocabulary construction. [1]

Finally, Blair and Maron allege that full-text systems, and STAIRS in particular, are not particularly user friendly in the sense that, in their test, even trained searchers were unable to achieve adequate performance, and untrained users would presumably do even worse.

Despite the impressive precision performance of the STAIRS system in the Blair and Maron test environment, the authors conclude with a surprising paraphrase of Samuel Johnson: "Full text searching is one of those things that . . . is never done well, and one is surprised to see it done at all!" ([1, p. 298]). This is surprising, moreover, because, in their study, no comparison was made between full-text-retrieval systems and manually indexed systems, nor between the retrieval performance of large versus small document collections. In this sense, conclusions drawn are unsupported by any data submitted to the reader—outside of the alleged poor recall performance exhibited by the STAIRS system in the legal case.

In fact, evidence abounds indicating that these conclusions may be more sentiment than fact. Specifically, the evidence from several retrieval evaluations conducted with very large document collections does not support the notion of output overload, although high recall naturally implies more retrieved items and hence more work in analyzing the output than low-recall searches. Moreover, comparisons between manual and automatic indexing systems on large document collections indicate that the automatic-text-based systems are at least competitive with, or even superior to, the systems based on intellectual indexing. Finally, there are automatic indexing systems that provide index terms that are not simply words extracted from document texts. Indeed, the automatic indexing results of Salton and Swanson [11, 20] that are cited in the Blair and Maron study were not based on the use of full document texts, but only on the analysis of document *abstracts*; the favorable results obtained in these studies on the effectiveness of automatic systems were achieved with abstracts (not full text), and therefore excessive input and verification demands were not placed on the system in these cases.

EXPERIMENTS WITH LARGE RETRIEVAL SYSTEMS

The Medlars Evaluation

In the late 1960s, Lancaster conducted an in-house study [7] of the Medlars demand search service, which is operated by the National Library of Medicine in Bethesda, Maryland, for searching biomedical literature. Medlars is based on manual, professional indexing by subject experts using a controlled indexing language described in the Mesh (Medical Subject Headings) thesaurus. After a manual indexing operation and a manual query formulation, the file search and retrieval operations are performed automatically.

The in-house evaluation of Medlars discussed in [7] involved searching a database of over 700,000 documents in biomedicine using a set of about 300 test queries. The search results varied widely; some queries performed perfectly (recall = 1.00, and precision = 1.00), whereas others retrieved no relevant

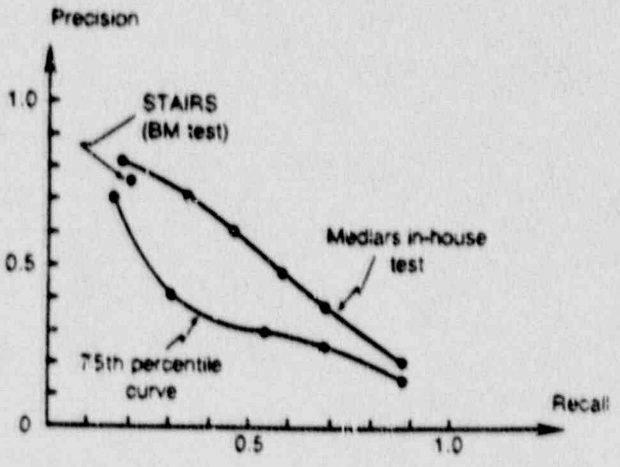


FIGURE 1. Medlars Search Service Evaluation (adapted from [7])

items at all (recall = 0, and precision = 0). For the 300 queries, the average recall performance was 0.58, and the average precision 0.50. In presenting the results, Lancaster notes that the actual performance value obtained for a query can be made to vary by submitting more or less specific query formulations. The average performance for a query can be made to slide along a monotonically decreasing curve starting at the high-precision/low-recall end of the performance spectrum, and proceeding to the high-recall/low-precision end as query formulations are broadened. The resulting curve representing the performance of the Medlars search system is shown in Figure 1: A second, lower curve (also included in Figure 1) represents the 75th percentile curve, giving the performance points exceeded for 75 percent of the test queries.

Three particular performance points for Medlars are analyzed in more detail in Table II. For the high-precision searches, the Medlars precision performance was about 0.80, but the recall reached only 0.19. For these searches, about 50 items were retrieved (out of some 700,000) of which about 40 were relevant. At the average performance point of 0.58 recall and 0.50 precision, the retrieved set increases to 175 documents of which about 60 percent were relevant on average. For high-recall searches, the recall reached nearly 90 percent (0.89), but the precision dropped to 0.20. To obtain that level of recall performance, it was necessary to retrieve between 500 and 600 items out of 700,000, of which about 130 on average were relevant to the query. Thus, the feared output overload predicted by Blair and Maron does not occur for the Medlars search service. This is most likely not due to the manual indexing but rather to the heterogeneity of the collections, which encompass all of biomedicine and would tend to facilitate the exclusion of useless material for any one search.

The set of 500 items retrieved on average for the Medlars high-recall searches represents only seven one hundredth of a percent (0.0007) of the collection; nonetheless, such a high recall entails substantial work for the users, and only specially motivated users (e.g., lawyers) might opt to submit such broad query formulations. In [7], Lancaster remarks that

> we can choose to operate Medlars, as it presently exists, at any performance point on or near the recall-precision plot (of Fig. 1) . . . . Intuitively one feels that Medlars should be operating at a higher average recall ratio (than 0.58) and should sacrifice some precision in order to attain improved recall. However Medlars is now retrieving an average of 175 citations per search in operating at recall 0.58 and precision 0.50. To operate at an average recall of 85 to 90 percent and an average precision of 20 to 25 percent implies that Medlars would need to retrieve an average of 500 to 600 citations per search. Are requesitors willing to scan this many citations to obtain a higher level of recall?

TABLE II. Medlars Performance Points

Performance points	Recall	Precision	Number of retrieved items	Number of relevant retrieved
High-precision searches	0.19	0.80	40–50	30–40
Medium performance	0.58	0.50	175	85
High-recall searches	0.89	0.20	500–600	135

By superimposing the performance point obtained in the Blair and Maron study of the STAIRS system—0.75 precision, 0.20 recall—on the Medlars performance curve in Figure 1, it can be seen that the STAIRS performance falls well within the range of the high-precision Medlars searches, even though no controlled language or manual indexing is used. The query broadening, recall-enhancing devices listed in Table I are available in an automatic environment like STAIRS just as they are in the Medlars controlled language environment.

The recall and precision failure analysis undertaken by Lancaster for the Medlars searches shows that manual indexing environments can also be problematic. A summary of the failure analysis for 797 recall failures (failures to retrieve relevant items) and 3038 precision failures (failures to reject nonrelevant items) appearing in Table III shows that a substantial proportion of the search failures are due to the manual indexing and the controlled language used in the Medlars environment. Some of these failures might be avoidable in an automatic indexing situation, whereas others would not. Poor search formulations and inadequate user–system interaction may occur with any retrieval system, manual or automatic. However, the conventional manual retrieval system is vulnerable in some very specific ways.

TABLE III. Typical Failures of Medlars Searches (adapted from [7])

Source of failure	797 recall failures (%)	3038 precision failures (%)
Indexing language (lack of appropriate term, false coordination)	10.2	36.0
Search formulation (too specific or too exhaustive)	35.0	32.4
Document indexing (too specific or too exhaustive)	37.4	12.9
Inadequate user–system interaction	25.0	16.6

Note: Some of the failures have multiple causes accounting for totals that may exceed 100 percent.

> if two people or groups of people construct a thesaurus in a given subject area, only 60 percent of the index terms may be common to both thesauruses:
>
> if two experienced indexers index a given document using a given thesaurus, only 30 percent of the index terms may be common to the two sets of terms:
>
> if two search intermediaries search the same question on the same database on the same host, only 40 percent of the output may be common to both searches:
>
> if two scientists or engineers are asked to judge the relevance of a given set of documents to a given question, the area of agreement may not exceed 60 percent. [3]

The solution Cleverdon offers is as follows:

> The problems caused by the use of a controlled language thesaurus and variations in (manual) indexing can be overcome by eliminating these two activities and using, as the input, an extract such as the title and abstract in natural (or free-text) language. Basically, a controlled language represents a reduction in the totality of the potentially available terms in the given subject area . . . (due to) compounding of real synonyms or spelling variations . . . (or to) subsuming of one or more specific terms by a general term . . . .
>
> Such combining of search terms as may, in a given search, be considered necessary is better done at the search stage than at the input. This appears to be one of the reasons why, in every test which has compared the performance of searching on controlled language index terms as against searching on abstracts in natural language, the results have been in favor of natural language. [3]

Comparison of Manual and Automatic Indexing

In the mid 1970s, a comparison between automatic and manual indexing was conducted using a NASA database consisting of documents from Scientific and Technical Aerospace Reports (STAR) and International Aerospace Abstracts (IAA). The test was based on a collection of 44,000 document titles and abstracts processed against 43 search requests. The following indexing systems were compared:

- a natural-language text-search system consisting of a machine search of document titles and abstracts, not the full text;
- a natural-language text-search system supplemented by a thesaurus of "associated concepts" prepared from the source documents;
- a controlled language indexing of the documents performed by human subject experts;
- the controlled indexing supplemented by natural-language terms extracted from the documents.

Title + Abstract here vs Full-Text Study (Blair ~~B~~ STAIRS study)

The search results for the NASA test as summarized in Table IV show that the natural-language abstract produces the best average recall for the 40 test queries (0.78) and also a high order of precision (0.63). The controlled language manual indexing produced a better precision value than the automatic abstract search (0.74), but a substantially worse recall (0.56). Based on these results, it is certainly not possible to conclude that searches of natural-language abstracts are inferior, in general, to controlled language indexing. Indeed, were the NASA search population legal personnel with a recall orientation similar to the searchers involved in the Blair and Maron test, they would certainly have preferred the output produced by the automatic search system with its recall advantages of over 20 percent compared to the manual system. Cleverdon, who was in charge of the NASA test, concludes that

TABLE IV. Comparative Evaluation of NASA Search System (adapted from [2]) (44,000 documents, 40 queries)

Indexing method	Recall	Precision
Natural-language indexing (text search of titles and abstracts)	0.78	0.63
Natural language supplemented by associated concepts	0.73	0.52
Controlled language manual indexing	0.56	0.74
Controlled language supplemented by natural-language terms	0.71	0.45

> within the parameters of this test, natural language searching on titles and abstracts proved at least equal to, and probably superior to, searching on controlled language terms; it also seems that a significant factor in this (result) was the increased level of indexing exhaustivity (provided by the natural language text search system). [2]

The performance points for the NASA search system evaluation are plotted in Figure 2, along with the curve representing the controlled term performance for the Medlars test, and an indication for the STAIRS system. Comparing NASA and STAIRS performance on collections of comparable size shows

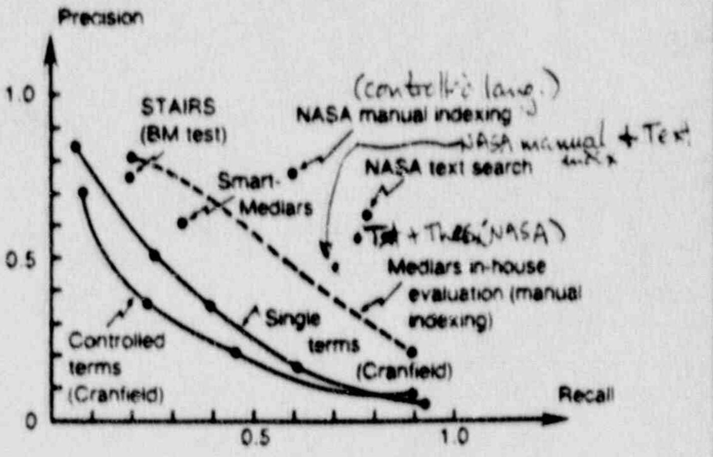


FIGURE 2. Comparison of Manual with Automatic Indexing

Cleverdon study showed high Recall + Precision using abstracts over text study #'s!

that the NASA searches are substantially more effective. Collection size does not seem to play an important role in search performance. Query type and homogeneity of subject matter are likely to be more important.

Many additional comparisons between automatic and controlled-term indexing systems appear in the literature. In [12], a small sample collection of 450 documents and 29 search requests is used to compare the performance of the Medlars system with an automatic indexing system based on abstract searching supplemented by the use of a thesaurus of related terms. The two systems produced almost identical results for the test collection: 0.31 recall and 0.61 precision for controlled-term indexing, versus 0.32 recall and 0.61 precision for natural-language terms plus thesaurus.

In the well-known Aslib–Cranfield study, an attempt was made to evaluate the performance of natural-language "single-term" indexing based on abstract searching and supplemented by many types of recall- and precision-enhancing devices. The automatically derived single-term languages were then compared with various kinds of controlled-term manual indexing systems [4] as applied against a sample collection of 1400 aeronautics documents tested by 221 queries. As shown by the two typical performance curves for the Cranfield study that are included in Figure 2 [4, pp. 127 and 164], the recall-precision performance for the Cranfield collection was relatively poor compared with other previously mentioned results obtained for much larger test collections. However, in practically every case, the Aslib–Cranfield tests indicate that the single-term natural-language indexing provided somewhat better search results than the comparable controlled-term indexing. This is true also for the two Cranfield searches illustrated on Figure 2.

However, as mentioned earlier, an automatic text-search system does not need to restrict itself to the use of single words extracted from document texts. Complete *automatic indexing* packages are available for constructing fairly sophisticated automatic document representations.

AUTOMATIC INDEXING THEORY AND PRACTICE

The effectiveness of any indexing system designed to produce useful content representations for written texts depends on two main characteristics: the *exhaustivity* of the indexing (i.e., the degree to which all aspects of the document content are recognized and represented in the indexed document representations), and the *specificity* of the individual index terms used to represent document content (i.e., the level of detail of a given content or index term). A high degree of exhaustivity tends to improve the recall performance of a search by permitting the identification of relevant materials that would remain unrecognized were the indexing exhaustivity lower, whereas a high degree of specificity is likely to favor search precision.

In principle, the choice of an indexing system that will be useful for content representation of natural-language texts should be based on linguistic considerations, especially semantic components. However, since linguistic analysis methods are difficult to apply efficiently to large text samples, most existing indexing theories are based on statistical or probabilistic methodologies. On the simplest level, both indexing exhaustivity and index term specificity may be characterized by the occurrence statistics of the terms in the collection of documents. In particular, the exhaustivity of the indexing is characterized to some extent by the number of index terms assigned to a given document, whereas term specificity is more or less inversely proportional to the number of documents to which a term is assigned [19]. Thus, terms that are assigned rarely may be assumed to be more specific than those more frequently assigned.

In judging the value of a term for purposes of content representation, two different statistical criteria come into consideration. A term appearing often in the text may be assumed to carry more importance for content representation than a more rarely occurring term, so that a document containing the term "pear" many times is likely to deal with the notion of pears. On the other hand, if that same term occurs as well in many other documents of the collection—that is, if all other documents also deal with pears—then the term "pear" may not be as valuable as other terms that occur more rarely in the remaining documents. This suggests that the specificity of a given term as applied to a given document can be measured by a combination of its frequency of occurrence inside that document (the *term frequency* or *tf*) and an inverse function of the number of documents in the collection to which it is assigned (the *inverse document frequency* or *idf*). The *idf* factor can be computed as 1 divided by the document frequency. A possible term weighting function for term i in document j [18] would then be

$$w_{ij} = tf_{ij} \times idf_i.$$

Using this term-importance definition, the best terms assigned to documents will be those occurring frequently inside particular documents but rarely on the outside. Such terms will in fact distinguish the documents of a collection from each other. Both factors of this equation are easy to calculate: The inverse document frequency of a term can be obtained in advance from a collection analysis, and term frequencies can be computed from the individual documents, as needed.

The Probabilistic Retrieval Model

In the *probabilistic retrieval* model, one assumes that the most valuable documents for retrieval purposes are those whose probability of relevance to a query is largest [10, 21]. The relevance properties of the documents can be estimated by using the relevance properties of the individual terms included in the documents. Under suitably simplified assumptions, a *term relevance* weight tr_i can then be generated for term i as

$$tr_i = \log \frac{N - n_i}{n_i} + \text{constants}$$

where N is the collection size and n_i represents the number of documents in the collection with term i [5]. This formula represents the importance of the *idf* factor, since the higher the document frequency n_i of a term, the lower the relevance weight tr_i. The probabilistic retrieval model thus provides some justification for the use of the *idf* factor in the term weighting formula given on page 653, since under appropriate mathematical assumptions the idf_i factor is approximately equal to the optimal probabilistic term weight tr_i.

The Term-Discrimination Model

A different but related way of approaching the document indexing task is basing the indexing on the *term-discrimination* model [18]. Under this model, it is assumed that the most useful terms for the content identification of natural-language texts are those best capable of distinguishing the documents of a collection from each other. This suggests that the value of a term should be measured by calculating the decrease in the "density" of the document collection that results when a given term is assigned to the collection. The density of the document space reflects the degree to which the document representations resemble each other. This density can be measured by computing the sum of the pairwise document similarities for all pairs of documents in the collection. This means that the density of the documents will be high when the documents resemble each other a great deal (i.e., when they are indexed by many of the same terms).

Using the term-discrimination approach, the broad, high-frequency terms become the least desirable content identifiers because they will be assigned to many documents in the collection, thereby enhancing the mutual similarity of the corresponding documents. The assignment of a broad high-frequency term, because it increases the average similarity between documents, also increases the document space density. If the discrimination value of a term is measured as the collection density before the given term assignment minus the density after term assignment, it is clear that high-frequency terms are characterized by a negative term-discrimination value. In the term-discrimination model, the very rare, low-frequency terms preferred by the *idf* factor are also not very desirable for content identification because they are assigned to so few documents that they hardly change the space density when introduced. The very rare terms thus receive a discrimination value close to zero.

The best content identifiers will be those occurring neither too rarely nor too frequently; they will be assigned to as many as one-tenth of the items in the collection and will serve to distinguish the items to which they are assigned from the remainder. A graphic representation of the variations in term-discrimination value as a function of the document frequency of terms is given in Figure 3. As the number of documents to which a term is assigned increases from zero, the term-discrimination value first increases from zero and becomes positive; then, as the document frequencies become still larger,

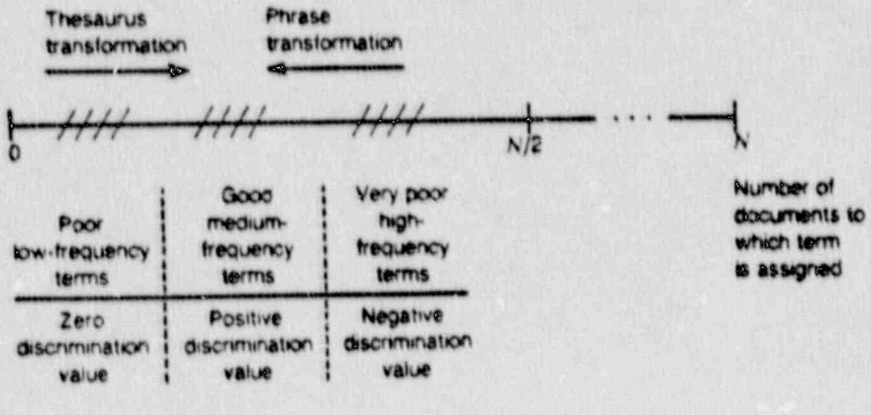

Thesaurus transformation
Phrase transformation
0
N/2
N
Number of documents to which term is assigned
Poor low-frequency terms
Good medium-frequency terms
Very poor high-frequency terms
Zero discrimination value
Positive discrimination value
Negative discrimination value

FIGURE 3. Term-Discrimination Model

term-discrimination values decrease rapidly and become negative for high-frequency terms.

The term-discrimination model confirms the notion that a correct degree of *specificity* exists for terms used as content identifiers, and that terms not exhibiting the appropriate specificity should be broadened when too specific or narrowed when too broad [19]. The recall- and precision-enhancing devices included in Table I can be used for this purpose. A principal method of term broadening involves using a thesaurus, or other vocabulary grouping device, to supply synonyms and related terms of various kinds to handle the text-independent relations between terms. Term narrowing is achieved by introducing term *phrases* to replace certain broad single terms, based on a text-dependent assessment. Under the term-discrimination model the thesaurus thus assumes a specific role as a grouping device for related narrow terms. Used in this way the thesaurus and phrase transformation methods produce shifts in terms toward the center of the frequency spectrum where the content identifiers with the best specificity are located.

A BLUEPRINT FOR AUTOMATIC INDEXING

These automatic indexing strategies make possible the design of effective automatic-text-based retrieval systems that are fully competitive with conventional manual operations and can be operated without the need for human subject or domain experts for document indexing and search formulation. Summarized below is a proposed basic pr [illegible] [13] for automatic indexing

- Identify the individual words occurring either in the documents or in document excerpts (e.g., titles and abstracts).
- Use a *stop list* of common function words (and, of, or, but, the, etc.) to delete from the texts the high-frequency funtion words that are insufficiently specific for content representation.
- Use a *suffix stripping* routine to reduce the remaining words to word stem form; this recall-enhancing transformation broadens the scope of the terms and can be performed automatically using a limited number of basic rules [9].
- For each remaining word stem *i* occurring in document *j*, compute a term weighting factor, which is the product of the term frequency of term *i* in document *j* multiplied by the inverse document frequency of term *i* in the collection as a whole. Available evaluation results indicate that term weighting improves retrieval effectiveness by distinguishing the important content terms from the less important ones [15].
- Represent each document by the chosen set of weighted word stems.

Retrieval evaluation results for this type of simple indexing for both large and small document collections indicate that even this *single-term* indexing method is competitive with, and often superior to, conventional intellectual indexing systems [2, 4, 12]. The STAIRS system used in the Blair and Maron test adheres to all these processes with the exception of term weighting. In STAIRS, term weights are assigned *after* retrieval of the documents based on term-occurrence characteristics in the retrieved document subset only; the weighting is then used to generate a *ranked* list of retrieved documents. The use of ranked document output improves the user-system interaction by alerting the user to the more important documents first; moreover, information culled from the documents retrieved early in the search can then be used to generate improved query formulations in subsequent searches.

Ideally, however, term weights should be generated before the query and document representations are compared during the search, and should be computed on the basis of the entire collection and not just a particular subset of retrieved items, which may or may not be representative of the entire collection. Certainly, terms exhibiting high-occurrence frequencies in the retrieved subset cannot be labeled effective or ineffective unless something is known a priori about their occurrence frequencies in the collection as a whole.

The basic indexing process can be improved by adding the following refinements:

- Generate weighted word stems that are attached to the documents.
- Use a thesaurus to replace terms with low document frequencies (and near zero discrimination values) by their corresponding thesaurus class identifications.
- Use a phrase-formation process to generate term phrases that incorporte terms with high document frequencies (and negative discrimination values) based on term cooccurrences in the document excerpts.
- Compute a combined term weight for assigned thesaurus classes and term phrases, and represent each document by the corresponding sets of weighted single terms, term phrases, and thesaurus classes.

In the STAIRS system, the thesaurus is generated "on the fly" by letting the user suggest terms that are synonymous, or related to, particular index terms. These related terms are then used automatically to expand the set of original terms. A previously available thesaurus that groups low-frequency terms into classes of related terms could be used for the same purpose.

A natural language query formulation can be converted into sets of weighted terms in the same way as a document text. Composite query-document similarity coefficients can then be computed, reflecting the similarities between corresponding term representations. When query-document similarity measurements are available, the documents can be ranked for output purposes in decreasing order of the query-document similarity. Moreover, improved query formulations can be generated by incorporating information obtained from the texts of previously retrieved documents [13].

When search requests are submitted in Boolean form, as they are in many operational retrieval environments, weighted terms can also be incorporated: Then, an approximate, fuzzy match between the weighted term sets representing the documents and the weighted Boolean query statements can be used to produce a query-document similarity measurement that is used in turn to obtain a ranked output in decreasing order of the query-document similarity. Term weighting and output ranking are therefore available for Boolean as well as non-Boolean queries [14, 17].

CONCLUSION

No support is found in the literature for the claim that text-based retrieval systems are inferior to conventional systems based on intellectual human input. Indeed, all the available evidence with reference to both large and small collections indicates that properly designed text-based systems are preferable to manually indexed systems. Furthermore, as Swanson pointed out over 25 years ago, ". . . it is expected that the relative superiority of machine text searching to conventional retrieval will become greater with subsequent experimentation as retrieval aids for text searching are improved, whereas no clear procedure is in evidence which will guarantee improvement of the conventional systems" [20].

REFERENCES

1. Blair, D.C., and Maron, M.E. An evaluation of retrieval effectiveness for a full-text document-retrieval system. *Commun. ACM 28*, 3 (Mar. 1985), 289–299. A recent evaluation of the IBM/STAIRS text-search system, which concludes that STAIRS does not always produce adequate search output.
2. Cleverdon, C.W. A computer evaluation of searching by controlled language and natural language in an experimental NASA data base. Rep. ESA 1/432, European Space Agency, Frascati, Italy, July 1977. A description of a large-scale test of the NASA search system using various manual and automatic text-analysis methods.
3. Cleverdon, C.W. Optimizing convenient on-line access to bibliographic databases. *Inf. Serv. Use 4* (1984), 37–47. A summary of the strengths and weaknesses of existing bibliographic retrieval systems and proposals for improving the existing methodologies.
4. Cleverdon, C.W., and Keen, E.M. *Aslib–Cranfield Research Project*. Vol. 2, *Test Results*. Cranfield Institute of Technology, Cranfield, England, 1966. The report on the most thorough evaluation of automatic versus manual text-analysis methods ever carried out, using a collection of 1400 aeronautics documents.
5. Croft, W.B., and Harper, D.J. Using probabilistic models of document retrieval without relevance information. *J. Doc. 35*, 4 (Dec. 1979), 285–295. Describes a method for using probabilistic considerations of term relevance for an initial collection search before any relevance information is available.
6. IBM World Trade Corporation. *Storage and Information Retrieval System (STAIRS)—General Information Manual*. 2nd ed. IBM Germany, Stuttgart, Germany, Apr. 1972. Contains an early description of the IBM/STAIRS system.
7. Lancaster, F.W. *Evaluation of the Medlars Demand Search Service*. National Library of Medicine, Bethesda, Md., Jan. 1968. An impressive description of the in-house test of the Medlars search system carried out at the National Library of Medicine.
8. Lancaster, F.W. *Information Retrieval Systems: Characteristics, Testing and Evaluation*. 2nd ed. Wiley, New York, 1979. A well-known textbook in information retrieval with an emphasis on system testing and evaluation.
9. Lovins, J.B. Development of a stemming algorithm. *Mech. Transl. Comput. Linguist. 11*, 1–2 (Mar. and June 1968), 11–31. A detailed description of an automatic word-stemming algorithm.
10. Roberison, S.E., and Sparck Jones, K. Relevance weighting of search terms. *J. ASIS 27*, 3 (May–June 1976), 129–146. Describes one of the main probabilistic information-retrieval models.
11. Salton, G. Automatic text analysis. *Science 168*, 3929 (Apr. 1970), 335–343. A survey of automatic text retrieval as of 1970.
12. Salton, G. Recent studies in automatic text analysis and document retrieval. *J. ACM 20*, 2 (Apr. 1973), 258–278. An evaluation of various automatic text-analysis and indexing methods.
13. Salton, G. A blueprint for automatic indexing. *ACM SIGIR Forum 16*, 2 (Fall 1981), 22–38. A relatively nontechnical summary of an approach to automatic indexing and text analysis.
14. Salton, G. A blueprint for automatic Boolean query processing. *ACM SIGIR Forum 17*, 2 (Fall 1982), 6–25. A summary of a retrieval system based on soft Boolean logic and automatically assigned term weights.
15. Salton, G., and Lesk, M.E. Computer evaluation of indexing and text processing. *J. ACM 15*, 1 (Jan. 1968), 8–36. An early set of test results for some automatic indexing methods.
16. Salton, G., and McGill, M.J. *Introduction to Modern Information Retrieval*. McGraw-Hill, New York, 1983. A recent textbook dealing with automatic text processing and text search and retrieval.
17. Salton, G., Fox, E.A., and Wu, H. Extended Boolean information retrieval. *Commun. ACM 26*, 11 (Nov. 1983), 1022–1036. A description of a retrieval model using soft (fuzzy) Boolean logic with weighted document terms and weighted Boolean queries.
18. Salton, G., Yang, C.S., and Yu, C.T. A theory of term importance in automatic text analysis. *J. ASIS 26*, 1 (Jan.–Feb. 1975), 33–44. Contains a description of term-discrimination theory and some retrieval results based on discrimination value weighting.
19. Sparck Jones, K. A statistical interpretation of term specificity and its application in retrieval. *J. Doc. 28*, 1 (Mar. 1972), 11–21. Relates the usefulness of index terms to certain statistical term occurrence parameters.
20. Swanson, D.R. Searching natural language text by computer. *Science 132*, 3434 (Oct. 1960), 1099–1104. A pioneering small-scale test comparing an automatic text-search system with a conventional retrieval system based on manual indexing; probably the earliest result showing the superiority of automatic text searching.
21. van Rijsbergen, C.J. *Information Retrieval*. 2nd ed. Butterworths, London, England, 1979. A well-known research-oriented information-retrieval text containing many original research results, including work in probabilistic information retrieval.

CR Categories and Subject Descriptors: H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing—*indexing methods, linguistic processing, thesauri*; H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval—*retrieval models, search process*; H.3.5 [**Information Storage and Retrieval**]: On-Line Information Services; I.2.7 [**Artificial Intelligence**]: Natural-Language Processing

General Terms: Design, Experimentation, Theory

Additional Key Words and Phrases: automatic text processing, automatic text retrieval, text analysis, text searching

Received 12/85; accepted 2/86

Author's Present Address: Gerard Salton, Dept. of Computer Science, Cornell University, 405 Upson Hall, Ithaca, NY 14853-7501.

ATTACHMENT #2 ATTACHED

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 "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 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, ① the construction of an indexing vocabulary, ② the training of indexers, and ③ 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 relevent.

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

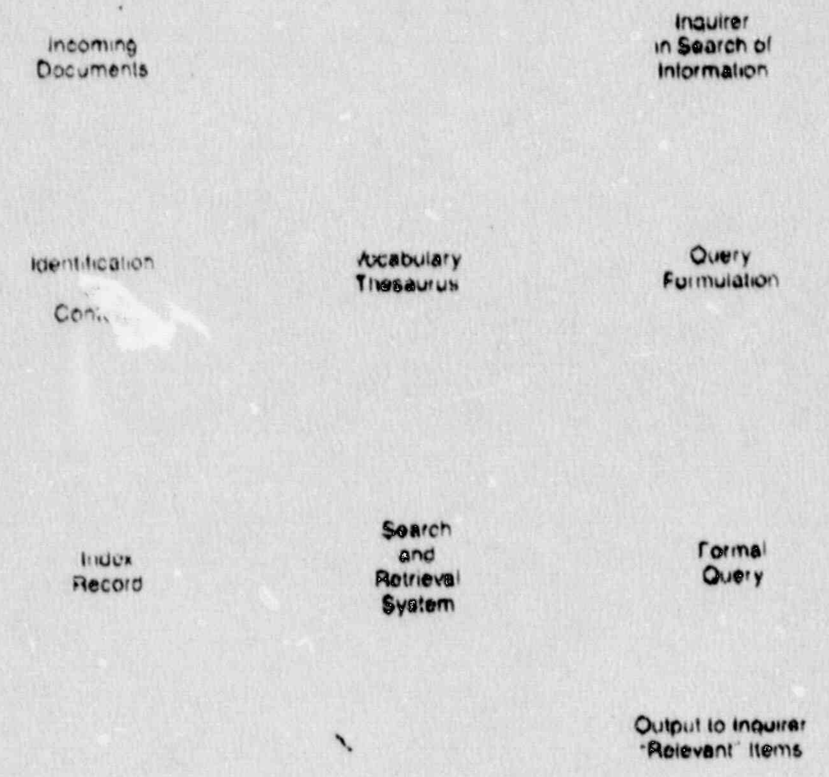

Incoming Documents
Inquirer in Search of Information
Identification of Content
Vocabulary Thesaurus
Query Formulation
Index Record
Search and Retrieval System
Formal Query
Output to Inquirer "Relevant" Items

FIGURE 1. The Dynamics of Information Retrieval

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

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

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 (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 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-

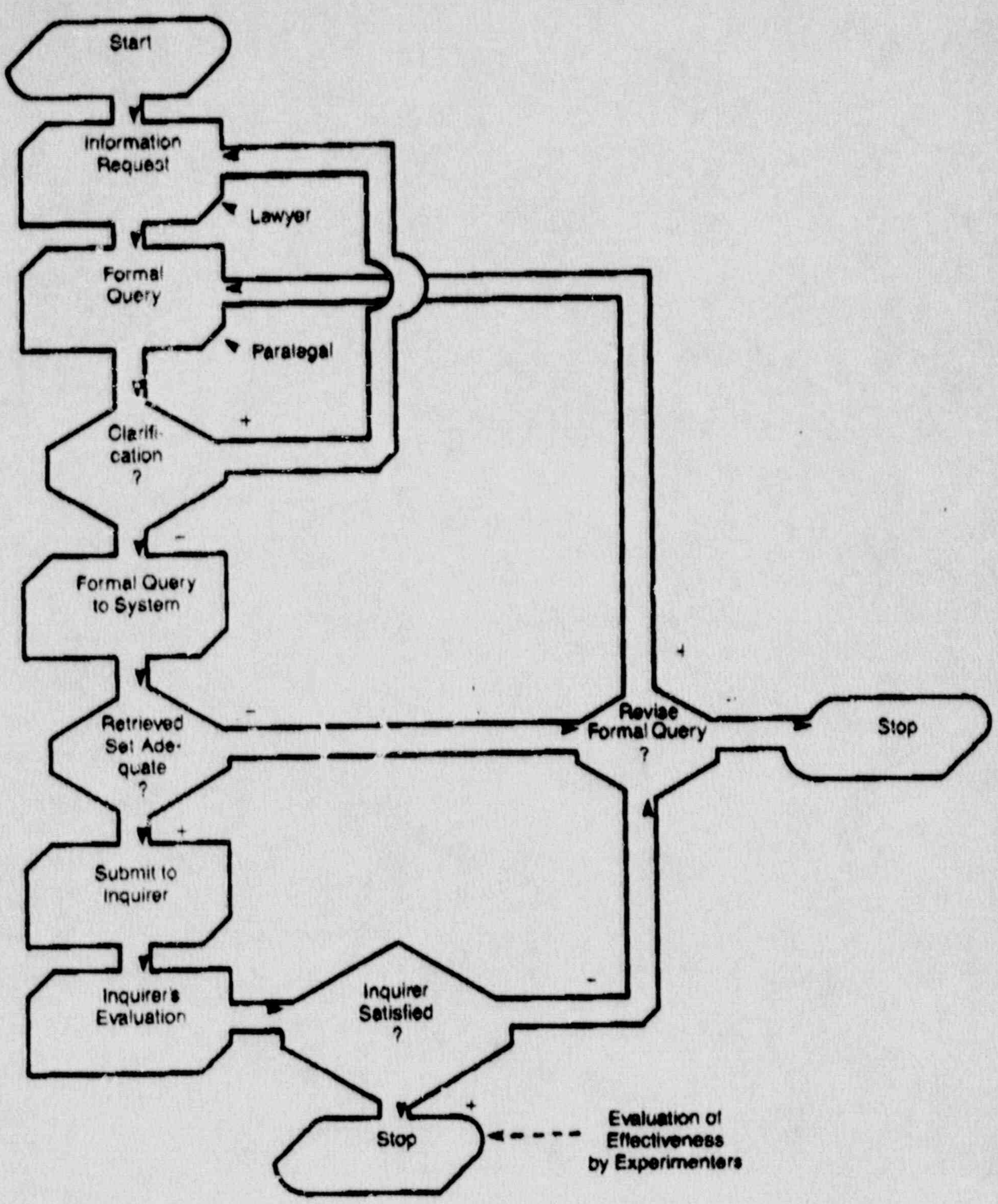


FIGURE 3. The Information Request Procedure

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 those 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
1	•	•
2	45.5%	92.6%
3	•	•
4	•	•
5	•	•
6	8.9	60.0
7	20.6	64.7
8	43.9	88.8
9	13.3	48.9
10	10.4	96.8
11	12.8	100.0
12	9.6	84.2
13	15.1	85.0
14	78.7	99.0
15	•	•
16	•	•
17	•	•
18	13.0	38.0
19	15.8	42.1
20	19.4	68.9
21	41.0	33.8
22	22.2	94.8
23	2.8	100.0
24	•	•
25	13.0	94.0
26	7.2	95.0
27	50.0%	42.8%
28	60.0	19.6
29	•	•
30	7.0	100.0
31	•	•
32	12.5	100.0
33	16.2	79.5
34	14.1	45.1
35	•	•
36	4.2	33.3
37	15.9	81.8
38	24.7	66.3
39	16.5	83.3
40	4.1	100.0
41	16.3	96.9
42	46.4	91.0
43	18.9	100.0
44	10.6	100.0
45	20.3	94.0
46	11.0	85.7
47	13.4	100.0
48	13.7	87.5
49	17.4	87.8
50	13.5	75.7
51	4.7	100.0

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

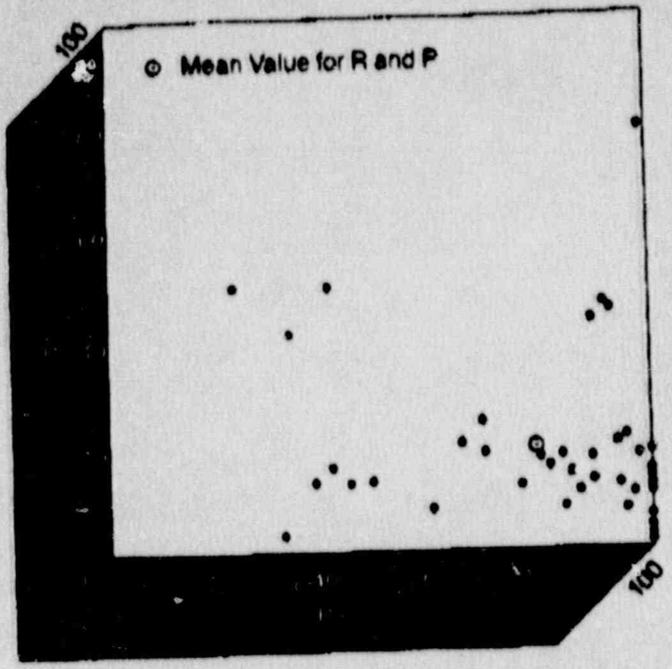


FIGURE 4. Plot of Precision versus Recall for All Information Requests

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 document in the entire file, and kr[illegible]g which documents the lawyer had previously j[illegible] relevant, we were able t cumpute 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 averag 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 t both the lawyer and the paralegal got 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 correspondence 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 [3]. 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 1960), 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, 1976.
3. Salton, G. Automatic text analysis. *Science 168*, 3920 (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 1960), 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 Precision

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.

ATTACHMENT #3 ATTACHED

ATTACHMENT #3

bottom line -
each method
-vocab vs abs - FT -
contribute
unique relevant
documents

CONTRIBUTIONS OF VALUE ADDED FIELDS AND FULL-TEXT SEARCHING IN FULL-TEXT DATABASES

Carol Tenopir, University of Hawaii at Manoa

Abstract: Some database producers assume that the availability of full text databases will make indexing and abstracting obsolete. Very few full text databases include both controlled vocabulary indexing terms and abstracts. As full text databases become more widely available, this assumption is beginning to be tested.

This study reviewed research to date that has examined full text retrieval performance on inverted file systems. Research comparing efficacy of searching on value-added fields vs. full text was also reviewed. Conclusions are not yet definitive but suggest that value-added field contribute to comprehensive retrieval and improve precision.

The author conducted a retrieval performance experiment in 1983-84 on the Harvard Business (HBR) full text database and the BRS search system. HBR contains controlled vocabulary descriptors and abstracts, allowing retrieval performance of these fields to be compared with full text.

Results showed that full text retrieved a high proportion of the relevant documents. Controlled vocabulary searching, and to a lesser degree abstracts, also contributed unique relevant documents. The value-added fields allowed much better precision in searching and had lower costs for searchers.

Unique relevant documents retrieved by each method were examined to judge the special contribution of each field Controlled vocabulary compensated for variations or changes in terminology, levels of specificity of terminology, and incomplete search strategy development. Abstracts pulled concepts together and somewhat standardized language. Full text allowed articles to be retrieved that contained relevant information peripheral to the article as a whole, compensated for deficiencies in controlled vocabulary, and often used more synonyms.

Suggestions for additional research will also be presented.

1. INTRODUCTION

Full text databases are increasing in numbers on the commercial inverted file search systems. Because full text databases are a relatively new phenomenon on the once traditionally bibliographic systems, much full text search strategy is based on assumptions or trial-and-error rather than on systematic study of the best results. Some producers or providers of full text databases assume that the availability and searchability of complete texts in inverted file systems will make indexing and abstracting obsolete. Few full text databases also include the value added fields of controlled vocabulary indexing terms and abstracts. This paper reviews some past research that compared search results and describes a recent project that examined the relative contributions of full text, controlled vocabulary terms, and abstracts in online search strategy on the Harvard Business Review database.

2. REVIEW OF RESEARCH

Many studies in information retrieval may be relevant to retrieval performance; described here are those few that examined full text vs. controlled vocabulary descriptors or abstracts on standard inverted file systems. The American Chemical Society (ACS) and BRS did a series of user studies of the full text of ACS journals before they were made commerically available on BRS. The researchers observed that searchers were able to find specific factual information by searching texts of articles when there were no corresponding terms in titles or abstracts [1].

Studies by Hersey et al. of the Smithsonian Institution Science Information Exchange (SSIE) compared retrieval performance from searching subject indexing codes with searching text words in a database of one-page summaries of research in progress [2]. An early version of the Mead Data Central software was used. The study concluded that retrieval performance with indexing terms was superior to that when searching free text words. Recall was about 30% higher; precision was 15-20% better. Both approaches offered advantages by retrieving documents that were unique and relevant. Text word retrieval provided detail; index code retrieval retrieved concepts and broad subjects and contributed to more complete retrieval. The authors recommended combination systems rather than forcing searchers to rely on one search technique.

Indexing may be expensive from the database producer's viewpoint, but free text searching can be expensive to users in terms of computer time. Three times as much computer time was required for the free text searching in the SSIE study as for the controlled indexing searching. The free text searches required three and one-half times the number of terms per question, and 14 times as many term combinations. In a study by Stein et al., six expert patents classifiers were asked to conduct 12 patent searches each on a LEXIS database of 50,000 patents. After the searches were completed, each query was studied to determine where in each patent the search terms occurred and what term variants occurred [3]. Results indicated that the full text resulted in substantially better retrieval than any single patent representation. Combinations of document segments were ranked by how often the full search query would be retrieved if a search was limited to them. A combination of summary and description provided the best search results (87%), followed by title and claims (16%) and title and abstract (7.5%). When individual segments were examined, full text, summary and description were of most help for retrieval while titles, abstracts, and claims were of limited help.

Several conclusions are suggested by these studies. No one method of searching (e.g. full text, abstract, controlled vocabulary descriptors, title) provides total recall in standard search systems and no one method consistently provides the best results. Controlled vocabulary searching, abstract, and full text searching retrieve unique documents, suggesting that the best strategy is to use a combination of methods.

3. HARVARD BUSINESS REVIEW STUDY

In an experiment conducted in 1983-1984, the author compared results from searching on words in complete texts, abstracts, and controlled vocabulary descriptors using the Harvard Business Review (HBR) full text database on the BRS search system. HBR has both controlled vocabulary descriptors and abstracts in addition to the complete texts of every article from 1976 to the present.

In a series of 31 questions, the text achieved on the average a relative recall of 73.9%. Controlled vocabulary had an average relative recall of 28% and abstracts 19.3%. Full text had the poorest average precision ratio of the three — 18% as compared to 34% for controlled vocabulary and 35.6% for abstracts. Full text searching was the most expensive with an average unit cost per relevant document of $7.86, as compared to $3.54 for controlled vocabulary searching and $3.89 for abstract word searching.

Although the full text contributed the highest recall, each of the three search methods contributed unique relevant documents in different questions. No one search method consistently provided all relevant documents. In only nine of the 23 topics that retrieved relevant documents were all documents retrieved by the full text. For the rest of the 23 topics the abstract and/or controlled vocabulary were required to achieve comprehensive retrieval. Samples of relevant documents that were retrieved by only one search method were examined in an attempt to characterize the unique contribution of each representation. This characterization may assist searchers to decide which search method would be best for a given topic.

3.1 Controlled Vocabulary

In nine questions relevant documents were retrieved by the controlled vocabulary that were not retrieved by any other search method. After examining these documents, there seem to be three major reasons why the controlled vocabulary resulted in retrieval while the full text did not. These reasons are: 1) variations or changes in terminology, 2) specificity of terminology, and 3) incomplete search strategy development by the searcher.

Terminology used in the texts of the articles in question varied from the more commonly used terminology found in similar articles. A relevant article retrieved only by controlled vocabulary in one question, for example, was a reprint of an article originally published in 1950. In 1950 the now common terms of "product diffusion" or "early adopters" were not in use. HBR's controlled vocabulary retained the older term "new products", use of which in the search strategy would have retrieved the 1950 reprint. In another question the controlled vocabulary term "family" retrieved a document relevant to personnel policies for spouses working in the the same firm. Nowhere in this document were the terms nepotism, couples, marriage, married, or spouse found, but the terms wives, wife, or relatives would have resulted in retrieval by the full text. The author of the document assumed male-owned firms that were hiring relatives (including wives); the searcher failed to add the appropriate synonyms. In another question a relevant document retrieved by the descriptor "flexible working hours" was not retrieved by the full text search because only the term "flexitime" was used in the text of the article. The searcher used the alternate spelling "flextime", but failed to use "flexitime."

These three questions point out the need to use both modern and older forms of words and to use many synonyms to achieve complete full text retrieval. The constancy of controlled vocabulary terms for any concept as compared to the inconsistent and changing nature of text language often assists retrieval.

Other relevant articles retrieved only by controlled vocabulary were retrieved because the controlled vocabulary terms were much broader than the subject requested or because there were terms for only one of two facets of the question. In one, for example, the user requested documents on the retirement of farmers and ranchers. Both concepts were specified in the full text and abstract searches. The HBR controlled vocabulary does not contain a term for the farmers or ranchers concept, however, so the single broad term "retirement" was searched. Some aspects of retirement planning are independent of the retiree's occupation, however, so some relevant documents were found with the broader strategy.

One question asked for articles on in-plant recreational facilities. Again, a fairly specific strategy using both concepts was conducted for the full text and abstract searches. Only the very broad term "employee benefits" was available for searching in the controlled vocabulary. The additional relevant documents retrieved by this strategy used the term "perks" rather than benefits in the text. The recreational facet was represented by such terms as relaxation, entertainment or recreational facilities. Another question also contained two concepts, only one of which was available in the controlled vocabulary. In a question on in-house databases, "Information systems" or "databases" retrieved relevant items that were not retrieved by the full text for two reasons. The first reason is that only the terms computer or data processing were used throughout the texts. The other reason is that the second facet of "office" or "inhouse" excluded relevant articles. The authors of the articles assumed any computer system or database was located "inhouse" or in the office without explicitly using those terms.

In summary, the strength of controlled vocabulary to control synonyms and varied or changing vocabulary was supported in this study. In full text searching on the standard commercial systems such as BRS the burden of compensating for language inconsistency is on the searcher. Controlled vocabulary costs the database producer more to create, but retrieves items difficult to find using full text only. Ironically, the limitations of a broad controlled vocabulary contributed to more complete retrieval without achieving unacceptable precision when no terms were available for both concepts of a search, because the single broader concept retrieved relevant items without adversely affecting precision. In a larger database this might cause unacceptable precision levels.

3.2 Abstract

The abstracts did not contribute as many unique relevant documents as did the controlled vocabulary. There was high overlap of abstracts with full text, which in a way shows the success of the HBR abstracters in summarizing the content of each article in the author's own words. Still, the relevant documents retrieved only by abstracters were examined to determine why they were not retrieved by any other method.

There seem to be three main reasons why relevant articles were retrieved by abstract searching but not by full text. These reasons are: 1) words did not appear in the same text paragraphs, 2) language varies in texts, and 3) the searcher did not use all possible synonyms in the search strategy.

The most common reason for abstract-only retrieval resulted from using the SAME paragraph operator in the full text searches instead of the broader Boolean AND. This decision was made because BRS and HBR both recommend limiting full text searches to the same paragraph. Search terms from both facets of a search appeared somewhere in many of the texts of these documents but the terms did not appear in the same grammatical paragraphs. In the abstract the important concepts were brought together into the same field (i.e., paragraph). In a question about the effects of unions on the introduction of new technology, several articles were retrieved by the abstract because all of the ramifications of unions were listed in the abstracts. The texts discussed each of these effects in turn without repeating the term "union". The same is true in a question about second careers. In one article retrieved only by abstract words, the concept of training or retraining was not mentioned in the same text paragraphs as the concept of new jobs or layoffs, but these two concepts were brought together in the abstract. For a question on stress of working wives the article that was retrieved only by the abstract is about the stressful role of corporate wives. A mention of them entering the workforce was in a paragraph without other search terms, but all concepts were together in the abstract field.

Another reason for document retrieval by the abstract but not by the full text is one of language. In articles uniquely retrieved by the abstract in one question, "wives" or "wife" are used more frequently in the text than "woman" or "women". The abstract uses women. If the synonyms "wives" or "wife" had been added to the full text search, this article would have been retrieved. In another question an article about Sioux Indians does not refer to them as a "minority" group in the text, but the abstract uses this term.

As with controlled vocabulary, it appears that the abstracts in HBRO sometimes compensate for the inconsistency of language and the necessity of many possible specific terms for the same concept. A comprehensive full text search requires listing many synonyms for each concept.

3.3 Full Text

Full text searching often retrieved many more unique relevant documents than either controlled vocabulary or abstract searching. One frequent contribution of the availability of full texts is, thus, an increase in the number of documents retrieved. An examination of a portion of the relevant documents retrieved only by the full text revealed four major characteristics. These are: 1) level of specificity can better match the question, 2) full text can compensate for deficiencies in the controlled vocabulary, 3) some concepts that are implied in the abstract but not mentioned explicitly are mentioned in the text, 4) full text sometimes uses more synonyms and can thus compensate for incomplete search strategies.

Articles that on the whole are broader in scope than the search request (that include the search topic as only a minor portion of the article) are the major reason for full text-only contributions. The abstracters and indexers attempt to match the depth or level of specificity of each article taken as a whole. Thus, an article on unionization of professional employees may list the specific professions in the text, but these are not mentioned in the abstract or controlled vocabulary terms. For documents retrieved only by abstracts the opposite was sometimes found—terms in the abstract were broader than the text terms. In a specific question about the effect of labor unions on the decline of productivity in the U.S., some articles mentioned many reasons for this decline, including labor unions. The specific reasons are accessible only via the full text where they are listed or mentioned briefly. This variance in the level of specificity was the one major reason for many of the text-only retrievals.

Another contribution of the full text is that it compensates for deficiencies in the controlled vocabulary. Several topics did not have appropriate descriptors for a concept, so narrower or broader terms had to be used. One question was about collective bargaining in colleges and universities but there are no HBR descriptors for colleges or universities. The "product and service" terms were used, but relevant articles discussed colleges and universities as a subject, not as a product or service. The same reason applies to a question about collective bargaining in libraries, schools, etc. HBR's policy of assigning only five descriptors means that only the major issues in an article are indexed. This, plus the policy of indexing and abstracting at the leve. of specificity of the article as a whole, results in many full text-only retrievals. All articles retrieved by the full text only seemed to have appropriate index terms within the constraints of the controlled vocabulary and the HBR indexing policy, however.

Compared to abstracts, full text facilitates retrieval of articles that mention a specific facet of a topic, but that are generally broader in scope than the search question. Full text also retrieved some articles when one facet was assumed but not explicitly mentioned in the abstract. For example, in the question about recreational facilities as benefits in organizations, the abstracts of some documents implied that recreational facilities provided to employees to reduce tension are benefits, but the term "benefit" was not explicitly used. In a question about attitudes toward hard work the concept of "attitudes" or "feelings" about hard work was implied but not mentioned in the abstracts.

Abstracts sometimes used jargon or a single term for a concept in the text while the full text stated it in several ways. For example, in a question about minorites including Hispanics the title and abstract of an article referred to "Mexicans". The text, however, used various synonyms such as "hispanics", "chicanos", "Mexican-Americans", resulting in retrieval. In a question about layoffs or unemployment, "hard-to-employ" was the only term used in the abstract of one document to describe unemployed workers. Unemployment caused by layoffs was included in the article but the term layoff was found only in the full text.

3.4 Summary

This examination has analyzed the unique contribution towards comprehensive retrieval that is made by full texts, abstracts and controlled vocabulary searching. The controlled vocabulary indeed controls synonyms or language that changes over time. The abstract brings together major concepts in an article that may have been discussed separately in the text. It also somewhat standardizes language. The major contribution of the full text is made when an article is of broader scope than the search question or when one facet of a question is mentioned only as one possible factor in a broader issue. Specific terms or causes are often listed or discussed in textual paragraphs but are too minor or specific to be indexed or abstracted. Each search method makes its own contribution and often this contribution depends on the nature of the search question or the individual articles in the database. No one method is complete for every situation. Relevant articles will be missed and search costs may be higher if searchers do not have the option of choosing various methods of searching. Indexing and abstracting are not made obsolete in full text databases, all representations assist complete retrieval and provide their own unique contributions.

3.5 Suggestions For Future Research

Because full text databases have not been widely available on commerical search services for long, there has not yet been much research that examines their characteristics. The present study is thus only an early step in determining how full text databases might best be searched, but the conclusions must be limited to a relatively small database of the business literature. The methodology used in this study should be replicated in other subjects to see if retrieval performance and search results vary with the subject matter of the text and to see if low precision becomes an even greater problem in larger databases. Related research has indicated that language patterns vary with the nature of the discipline, but this has yet to be tested with full text online searching. Such an extension into other social science disciplines and into physical science disciplines could have important ramifications for searchers in search strategy development and for publishers in database design decisions.

Another variation on the present study would be to change the full text search strategies to use the Boolean AND operator rather than the paragraph SAME operator or to compare various full text strategies. This would help to identify the best full text strategy. The type of research mentioned so far is practical given the realities of the present systems, but can only suggest ways these existing systems might be improved. Any studies limited by the fundamental designs currently in use cannot reveal optimal performance in an ideal situation that has no previous design assumptions. Additional user studies are needed that will reveal how potential users would most like to use full text databases if they were not restricted by current system constraints.

Future research should take into consideration the different possible uses of full text, including browsing, fact retrieval, and finding articles on a given topic. Users with different types of needs may have different requirements for search and display features. The research on the use and retrieval characteristics of full text databases is just beginning.

NOTES

[1] Kay Durkin, et al., "An Experiment to Study the Online Use of a Full-Text Primary Journal Database," in Proceedings of the 4th International Online Information Meeting: 1980 December 9-11, London, England (Oxford, England: Learned Information, Ltd., 1980), pp. 53-56.

[2] David F. Hersey, et al., "Comparison of On-Line Retrieval Using Free Text Words and Scientist Indexing," in The Information Conscious Society: Proceedings of the American Society for Information Science 33rd Annual Meeting: 1970 October 11-15, Philadelphia, PA (Washington, DC: ASIS, 1970), pp. 265-268.

David F. Hersey, et al., "Free Text Word Retrieval and Scientist Indexing: Performance Profiles and Costs," Journal of Documentation 27 (September 1971):167-183.

[3] D. Stein, et al., "Full Text Online Patent Searching: Results of a USPTO Experiment," in Proceedings of the Online '82 Conference, 1982 November 1-3, Atlanta, GA (Weston, CT: Online Inc., pp. 289-294.

National ONLINE MEETING

PROCEEDINGS—1985

New York, April 30–May 2, 1985

Sponsored by
ONLINE REVIEW
The International Journal
of Online Information Systems

Compiled by
Martha E. Williams
Thomas H. Hogan

Learned Information, Inc.
Medford, NJ

ATTACHMENT #4 ATTACHED

An Evaluation of the Applicability of Ranking Algorithms to Improve the Effectiveness of Full-Text Retrieval. I. On the Effectiveness of Full-Text Retrieval*

Jung Soon Ro
56-35 Yukchon-2-dong, Eunpyung-ku, Seoul 122, Korea

It is generally accepted that information retrieval based on full texts of documents will result in higher recall and lower precision compared with retrieval using paragraphs, abstracts, or controlled vocabularies. Part I of the study tested this assumption by examining the effectiveness of full-text retrieval compared with other approaches in terms of recall and precision. Experiments were conducted on a subset of a journal-article collection with nine search questions through the BRS search service. Full-text retrieval was found to achieve significantly higher recall and lower precision than searches by other methods. Part II of the study will focus on how to improve the low precision of full-text retrieval without a decrease or with a minimum decrease in recall. Document-term-weighting algorithms proposed in past research for automatic extractive indexing were examined as a means to improve the low precision of full-text retrieval.

Introduction

A full-text retrieval system is a document-retrieval system in which the full tests of all documents in a collection are stored on a computer system so that every word in each sentence of every document can be located by computer software. Because of the continued decline in the cost of computer-storage devices and byproducts of computerized publication, full texts of documents are increasingly available in machine-readable form and serve as databases for information retrieval.

From previous research and commentary, it was expected that full-text retrieval would result in higher recall and lower precision when compared with retrieval using paragraphs, abstracts, or controlled vocabularies. The first purpose of this study was to test this assumption in journal articles by examining the effectiveness of full-text retrieval compared with that of other approaches in terms of recall and precision. The second purpose of the study was to examine how to improve the precision of full-text retrieval with minimum decrease in recall. Document-term-weighting algorithms proposed in past research for automatic extractive indexing were examined as a means to improve the low precision of full-text retrieval. Parts I and II of this study will address these two study areas in turn.

Background

Research on full-text retrieval started with studies on the possibility of automatic text analysis, i.e., the possibility of extracting keywords from full text of documents. In a relatively early study, Swanson reported the superiority of retrieval performance of a system based on automatic text analysis over a conventional system based on a manually assigned subject-heading index [2]. In later work, the SMART system predicted the superiority of full texts to abstracts in order to extract index words from them [3].

Many efforts to test the effectiveness of the full-text retrieval system have been made on portions of the legal literature [4–11]. Most of these studies reported the superior effectiveness of full-text retrieval compared with the manual, conventional technique of index lookup on court decisions. In the field of journal articles, some studies have been done on user-opinion survey for online vendor full-text systems [12–14] or on the usage of full-text searching for data retrieval [15–17]. Recently, Tenopir [18] examined the effectiveness of full-text retrieval where every word in a document is used for an index word, compared with that of other searches conducted on the fields of abstracts, titles, and controlled vocabularies. In Tenopir's study full-text retrieval was limited to paragraph searching, a kind of proximity searching which retrieves documents in which searching words appear within the same grammatical paragraphs.

*The research reported in this article is based on my doctoral dissertation, completed in December, 1985, at Indiana University, School of Library and Information Science [1].

Received April 21, 1986; accepted October 6, 1986.

Tenopir's Experiment

Tenopir's experimental data on the effectiveness of searches conducted on paragraphs, abstracts, and controlled vocabularies of journal articles was incorporated in this study. Tenopir tested the effectiveness of full-text retrieval (limited to paragraph searching) on the subset of *Harvard Business Review* (HBR) for the time period January 1976–August 1983, with 31 questions from the history files of two university libraries, through the BRS search system. Three faculty members of the business school judged relevance of documents retrieved from 31 questions. Paragraph searching in full texts of journal articles was found to achieve significantly higher recall and lower precision than searches by other methods. Also, paragraph searching was found to retrieve a significantly larger number of unique relevant documents than searches by other methods.

A use made of Tenopir's data was to investigate the reliability of relevance judgment for the present study. Since many documents are expected to be retrieved from full-text searching, and an enormous amount of time is required to judge documents for relevance, only one judge was used to assess relevance in the present study. However, since relevance judgment by one judge was questioned in the previous research [19], especially when questions came from a history file not directly from the judge, Tenopir's data on relevance judgment was used to investigate the reliability of relevance judgment assessed by one judge in this study. Also, using Tenopir's search strategy it was possible to be free from subjective bias for searching strategy which may affect the effectiveness of the retrieval.

Experimental Design

Database

The database used for the study was a subset of the *Harvard Business Review* for the time period of January 1979–August 1983. The descriptive characteristics of the collection of full texts of articles are given in Table 1.

Search Questions

Because of the enormous amount of time required to judge relevance of documents retrieved from each question, the search questions examined were limited to nine. In sampling nine questions, two factors were considered. First to study the difference between paragraph searching and full-text searching, at least one AND operator in search strategy is required, because in the paragraph search of Tenopir's study, the operator AND was replaced with SAME to retrieve a intersection "A AND B" within the same paragraphs. The second factor was considered for the second purpose of this study, which will be discussed in Part II of this article. In the experiments for the second purpose of this study, effectiveness of full-text retrieval was compared with that when weighting algorithms were applied to full-text retrieval. Tenopir used three operators for proximity searching: ADJ for adjacent words, WITH for words within the same sentences, and SAME for words within the same paragraphs. Since there is no obvious way to apply ranking algorithms to phrases, especially phrases of which single words appear anyplace in a sentence or a paragraph, only adjacent words were considered as phrases in applying weighting algorithms in this study. Thus, search questions stated with SAME or WITH for phrases were excluded in the sampling. By eliminating questions stated without AND operator or with a WITH or SAME operator for phrases, 20 out of 31 questions remained. From these, nine questions were selected at random. Table 2 lists the nine questions considered in this study.

TABLE 1. Descriptive statistics on full texts of articles.*

Total documents in the collection	448
Total words in the collection (collection length)	1,829,601
Number of word tokens (document length)	
Range	800–13,891
Mean word tokens per document	4,084

*Harvard Business Review, Jan./Feb. 1979–July/Aug. 1983.

Search Strategy

The same search strategy used in Tenopir's paragraph searching was conducted through the BRS search system for full-text search except for replacing the operator SAME with AND. Thus "A AND B" instead of "A SAME B" retrieved documents in which concepts A and B appear not only within the same paragraphs but also anyplace in a document. Table 3 shows the search strategy used for the nine questions.

Relevance Judgments

One doctoral student in the Department of Management, School of Business, Indiana University, assessed the relevance of documents retrieved both from full-text retrieval and from other retrieval methods conducted by Tenopir. Since nine questions are closest to the field of management rather than accounting, finance, or others in business, the judge was selected from the department. The knowledge of any Ph.D. student in the department was considered sufficient to judge the relevance of documents, since the HBR is a general and popular journal in the field of business. Recall and precision ratios associated with abstracts, descriptors, and paragraphs in Tenopir's study were recalculated using these new judgments. Although one judge has been found acceptable in previous research in the legal field of full-text retrieval in which questions were not submitted by the judge himself [20], the reliability of these judgments was investigated by computing the agreement of this judge with Tenopir's judges using both Holsti's coefficient of reliability (CR) [21] and Scott's index of reliability (*pi*) [22].

TABLE 2. Search topics.

1. I would like literature on cutback management or the process of transition management or administration.
2. Workaholism, workaholics, attitudes toward hard work.
3. Effect of diet and exercise programs on reduction of absenteeism and increase of productivity among corporate staffs and executives.
4. Scheduling of extended work hours. Computation of productivity and safety in relation to extended work hours. I am specifically interested in extended work hours over eight hours in relation to the above aspects.
5. Collective bargaining by women-dominated professions such as social workers, nurses, librarians, and teachers.
6. Impact of collective bargaining on the introduction of new technology.
7. Retirement planning by farmers or ranchers.
8. Productivity in Japan versus productivity in the US.
9. Productivity with unions versus productivity in nonunion companies.

Holsti's coefficient of reliability (CR) is a widely used coefficient of reliability, indicating the ratio of coding agreements to the total number of coding decisions. That is,

$$CR = 2M \ / \ M1 + M2,$$

where M is the number of coding decisions on which the two judges are in agreement, and $M1$ and $M2$ refer to the number of coding decisions made by judges 1 and 2, respectively. However Holsti's CR has been criticized because "it does not take into account the extent of intercoder agreement which may result from chance [21]." Scott's index of agreement between two coders, i.e., $pi = Po - Pe \ / \ 1 - Pe$, takes into account both the observed proportion of agreement (Po) and the porportion that would be expected by chance (Pe). Compared with the agreement between three judges in Tenopir's study (CR = 66%), the agreement between this judge and Tenopir's judges was 87%. When considering the extent of interjudge agreement which may result from chance, Scott's index of reliability pi was 71%.

Findings

As shown in Table 4, on the average, based on the nine questions studied, the full-text approach retrieved 68 documents, while paragraph searching retrieved 17.1 documents, abstract retrieved three documents, and controlled vocabulary retrieved 1.7 documents. Compared with Tenopir's experiments retrieving 17.8 documents from paragraph, 2.4 from abstract, and 3.2 from controlled vocabulary, with 31 questions in the database of January 1976–August 1983, abstract search retrieved more documents than did controlled-vocabulary search. The reason abstract search retrieved more documents than controlled vocabulary in this study seemed to be because questions 23, 31, and 35 were not considered in this study. In Tenopir's study for those three questions, broad search techniques were used for natural-language searches. For example, for question 31, the broad concept "information system" was searched in the field of controlled vocabulary, while a specific concept "personal information system" was used for abstract search by intersecting the two concepts "information system" and "personal." Questions 31 and 35 were not considered in this study because of operators SAME or WITH, and question 23 was not selected at random. Table 5 contains the number of documents retrieved from each search.

TABLE 3. Search strategies.

Question		
1.	1.	cutback$ or cut ADJ back$ OR transition
	2.	manage$ OR administ$
	3.	1 AND 2
2.	1.	workahol$
	2.	hard ADJ work
	3.	attitude$ OR belief$ OR feeling$ OR believ$
	4.	1 OR (2 AND 3)
3.	1.	absentee$ OR productivity OR motivation
	2.	executive$ OR employee$ OR worker$ OR personnel
	3.	diet OR exercise OR health OR nutrition OR physical ADJ fitness
	4.	1 AND 2 AND 3
4.	1.	flexible ADJ hour$ OR flex ADJ time OR flextime OR overtime OR four ADJ day ADJ week OR extended ADJ hour$
	2.	schedul$ OR productivity OR safety
	3.	1 AND 2
5.	1.	librar$ OR nurse$ OR hospital$ OR social ADJ work$ OR teacher$ OR educator$
	2.	woman OR women OR female
	3.	Occupation$ OR job OR jobs OR profession$
	4.	strike OR strikes OR union$ OR collective ADJ bargain$ OR negotiat$
	5.	2 AND 3
	6.	(1 OR 5) AND 4
6.	1.	collective ADJ bargain$ OR strike$ OR union$ OR negotiat$
	2.	technolog$ OR automat$ OR robot$ OR computer$ OR minicomputer$ OR microcomputer$ OR mechanization
	3.	1 AND 2
7.	1.	farm$ OR ranch$
	2.	retire$
	3.	1 AND 2
8.	1.	productivity
	2.	japan
	3.	united ADJ states OR us OR u ADJ s OR america$
	4.	1 AND 2 AND 3
9.	1.	productivity
	2.	union$ OR collective ADJ bargain$ OR strike OR strikes OR open ADJ shop$ OR nonunion OR non ADJ union
	3.	1 AND 2

TABLE 4. Average number of documents and relevant documents retrieved in Tenopir's study and this study.

	Full text	Paragraph	Abstract	Control
Tenopir's Study				
No. of documents retrieved		17.8	2.4	3.2
No. of relevant documents retrieved		3.5	1	1.2
This Study				
No. of documents retrieved	68	17.1	3	1.7
No. of relevant documents retrieved	8.4	5	1.8	1

TABLE 5. Number of retrieved documents from each search.

Question	Full text	Paragraph	Abstract	Controlled
1	86	40	3	6
2	20	5	2	0
3	91	6	0	0
4	18	5	2	2
5	102	18	2	0
6	177	49	9	0
7	15	0	0	3
8	31	7	3	3
9	72	24	6	1
Total	612	154	27	15
Mean	68	17.1	3	1.7

TABLE 6. Number of relevant documents from each search.

Question	Union	Full text	Paragraph	Abstract	Controlled
	29	29	16	3	4
	4	4	3	0	0
3	1	1	1	0	0
4	3	2	2	1	2
5	2	2	2	1	0
6	14	13	6	4	0
7	1	0	0	0	1
8	11	11	5	3	1
9	15	14	10	4	1
Total	80	76	45	16	9
Mean	8.9	8.4	5	1.8	1

The number of relevant documents retrieved from each search is given in Table 6. On the average of nine questions, out of a total number of 8.9 relevant documents, full-text search retrieved 8.4 relevant documents, while paragraph search retrieved five documents, abstract retrieved 1.8, and controlled vocabulary retrieved one document. Full-text search retrieved 1.7 times more relevant documents than paragraph search, 4.7 times more than abstracts, and 8.4 times more than controlled vocabularies. In question 7 a relevant document retrieved by the descriptor "flexible working hours" was not retrieved by both paragraph and full text search because only the term "flexitime" was used in the text of the article. As mentioned by Tenopir, she used the alternate spelling "flextime," but failed to use "flexitime."

The relevant documents retrieved from full texts or paragraphs had lower relevance value than that from controlled vocabularies or abstracts. Table 7 shows that the relevance degree of relevant documents retrieved by full text is an average of 3.566, compared with 3.644 for paragraphs, 3.875 for abstracts, and four for descriptors, when the weight of four is assigned to the documents judged "definitely relevant" to questions and the weight of three is assigned to the documents judged "probably relevant." All relevant documents retrieved by controlled vocabularies were definitely relevant.

Tables 8 and 9 translate these numbers of relevant documents retrieved from each search to relative recall and precision ratios. Relative recall was substituted for recall and defined as the number of relevant documents retrieved by a single search divided by the number of relevant documents in the union of sets retrieved by several searches on the same topic. Recall and precision ratios in Tables 8 and 9 are macroaveraged, in which a parameter (recall or precision) is calculated for each question and the average is then taken. Compared with 61.31% for paragraphs, 18.37% for abstracts, and 21.49% for controlled vocabularies, full texts rated 83.65% of recall. On the other hand, full text achieved lowest precision, an average of 14.46%, compared with 36.64% for paragraphs, 58.73% for abstracts, and 66.67% for controlled vocabularies. From the microaveraging viewpoint which totals over the set of questions, full text achieved 95% of recall compared with 56.2% for paragraphs, 20% for abstracts, and 11.2% for controlled vocabularies. The precision of full-text searching, by

TABLE 7. Relevance degree of relevant documents retrieved from each search.

	Full text	Paragraph	Abstract	Controlled
No. of relevant documents retrieved	76	45	16	9
No. of documents judged definitely relevant	43	29	14	9
No. of documents judged probably relevant	33	16	2	0
Mean	3.566	3.644	3.875	4

TABLE 8. Relative recall ratio of each search.

Question	Full text	Paragraph	Abstract	Controlled
1	100.00	55.17	10.35	13.79
2	100.00	75.00	0.00	0.00
3	100.00	100.00	0.00	0.00
4	66.67	66.67	33.33	66.67
5	100.00	100.00	50.00	0.00
6	92.86	42.86	26.67	0.00
7	0.00	0.00	0.00	100.00
8	100.00	45.45	20.00	6.67
9	93.33	66.67	25.00	6.25
Mean	83.65	61.31	18.37	21.49

TABLE 9. Precision ratio of each search.

Question	Full text	Paragraph	Abstract	Controlled
1	33.72	40.00	100.00	67.67
2	20.00	60.00	0	···
3	1.10	16.67	···	···
4	11.11	40.00	50.00	100.00
5	1.96	11.11	50.00	···
6	7.34	12.24	44.44	···
7	0	···	···	33.33
8	35.48	71.43	100.00	33.33
9	19.44	41.67	66.67	100.00
Mean	14.46	36.64	58.73	66.67

microaveraging, was 12.42% compared with 29.22% for paragraphs, 59.26% for abstracts, and 60% for controlled vocabularies.

Discussion of the High Recall and Low Precision of Full-Text Retrieval

Besides the two pioneering works which predicted the possibility of full-text document retrieval [2, 3], much research reported the more relevant documents retrieved by full-text searching compared with other searches. The Smithsonion Institution Science Information Exchange Project reported the 30%–40% higher average recall value of full-text searching ~~than~~ that of a search using subject-indexing codes, although the database consisted of one-page summaries of a research project has many of the characteristics of a lengthy abstract database rather than full text [4]. In the legal field, the LITE system found that in 7.5% of the total searches full-text searching retrieved fewer relevant citations than were discovered by a manual, conventional technique of index-lookup search, while in 48.4% full-text searching retrieved more relevant citations than were discovered manually [5]. The Joint American Bar Foundation and International Business Machine Project reported that the full-text search and the manual search had performed about equally well in terms of recall, and that the manual search was about twice as effective in terms of precision [6]. The Oxford experiment found that full-text search had 70% of recall and 29% of precision, while manual index-lookup search had 49% of recall and 92% of precision [7]. The Responsa Project found that the average precision of full-text search was 34% when recall was achieved up to 100% for all questions [8]. However, the optimism of a full-text document retrieval system was not supported by everyone. Blair and Maron evaluated the IBM's full-text retrieval system, STAIRS, with a legal-document collection, and reported low recall and high precision, i.e., a recall value of 20% and a precision value of 79% [11].

Wrong! Ans. to FW Lancaster in Letter to Editor JASIS Sept '89 p 362

The contrast of the effectiveness between Blair and Maron's and other studies' has theoretical reasons. The first is the different definition of recall value used in the research. To count the total number of relevant documents (including unretrieved), Blair and Maron sampled a subset of document collection, examined the samples to assess the relevant documents, and then estimated the total number of relevant documents. On the other hand, other studies, including this study, used the relative recall value by which the total number of relevant documents was defined as the number of relevant documents in the union of sets retrieved by several searches on the same topic. Since only the documents retrieved by search methods operated, rather than the entire database or a sample size of the database, are examined to assess the relevance judgment, the total number of relevant documents in the union of sets retrieved by searches operated might be fewer than (or at greatest equal to) that in the entire database or a sample size of a database (of course, under random sampling). It is therefore possible that the recall value of a search system is higher when only the documents retrieved by search methods operated are examined than when the entire database or a sample size of databases is examined for relevance judgment, under the same conditions of other variables.

The second reason for the two contradicting results on the effectiveness of full-text searching relates to the search strategy and the inverse relationship between recall and precision, where recall often increases as precision decreases, and vice versa. As mentioned by Blair and Maron, adding intersecting terms to a query results in narrowing the size of the output, while adding alternative terms (like synonyms) with the Boolean OR operator increases the size of the output. In other words, the possibility of retrieving relevant documents decreases with the increased number of intersecting terms, but increases with the increased number of alternative terms per concept. The process of adding intersecting terms to a query, continuing until the size of the output reaches a manageable number, was reported to be necessary because of the large size of database in Blair and Maron's study. Search queries employing four or five intersecting terms were not uncommon among the queries used in Blair and Maron's test.

The too many missed additional terms, as mentioned by Blair and Maron, which also resulted in the low recall and high precision in their study, might be related to the characteristics of legal documents. Blair and Maron reported that 20% of relevant documents were retrieved by full text. By examining the other 80% relevant documents unretrieved, they found too many additional terms, up to 26 or 44 other words, to retrieve relevant documents that had been missed. The characteristics of legal documents resulting in too many missed additional terms were examined in their report, which was not possible in journal collections. Using Tenopir's search strategy with journal articles, this study did not manipulate the search output. Compared with the up to more than 44 alternative words and the four or five intersecting terms commonly used for a query in Blair and Maron's test, only 3.4 alternative words on the average of nine questions and two, or at most three, intersecting terms were used in this study, in which the 95% of relevant documents were retrieved from full-text searching. Journal articles seemed to have fewer alternative words than did legal documents; these characteristics might affect the effectiveness of full-text searching.

All the missed words were used to retrieve all the relevant document by full-text searching in Blair and Maron's study, i.e., with 100% of recall, what would the precision value of it be? The Responsa Project suggested the possible answer, i.e., 34% of precision.

Conclusion

As predicted, there was a significant difference between the full-text search and searches by other methods. The hypothesis stated as a null was rejected using ANOVA at the significance level of 0.0001 for recall and at the level

0.0032 for precision ratio. That is, full-text retrieval resulted in significantly higher recall and lower precision than searches by other methods. Scheffee and Tuckey HSD showed that full-text search significantly differed from abstract search and controlled vocabulary search, but not significantly from paragraph search, for both recall and precision. The relevant documents retrieved from full-text searching were judged less relevant to questions than that from other searches.

Since full-text searching retrieved the greatest number of relevant documents, it is recommended for a comprehensive search if cost effectiveness is not required. However, the relevant documents retrieved by controlled vocabularies or abstracts had more relevance value; controlled-vocabulary searching or abstract searching is recommended for brief searching.

Because research on full-text retrieval has just started, further investigation could be fruitfully conducted in many related areas. Subject matter of databases, size of databases and questions, search strategies, types of documents (journals, encyclopedias, newspaper, etc.), etc. could be studied as the factors affecting the effectiveness of full-text retrieval for further studies. Research on a possibility to improve the precision of full-text retrieval will be presented in Part II of this article.

References

1. Ro, J. S. "An Evaluation of the Applicability of Ranking Algorithms to Improving the Effectiveness of Full Text Retrieval." Ph.D. dissertation, Indiana University, 1985.
2. Swanson, D. "Searching Natural Language Text by Computer." *Science*. 132:1099–1104; 1960.
3. Salton, G.; Lesk, M. E. "Computer Evaluation of Indexing and Text Processing." *Journal of the Association of Computing Machinery*. 25:8–36; 1968.
4. Hersey, D. F.; Foster, W. R.; Stalder, E. W.; Carlson, W. T. "Free Text Word Retrieval and Scientist Indexing: Performance Profiles and Costs." *Journal of Documentation*. 27:167–183; 1971.
5. David, R. P. "The LITE System." *Judge Advocate General Law Review*. 8:6–10; 1966.
6. Eldridge, W. B. "An Appraisal of a Case Law Retrieval Project." In: D. Johnston, Ed. *Computer and the Law*. Kingston; 1968.
7. Tapper, C. *Computer and the Law*. London: Weidenfeld and Nicolson; 1973.
8. Schreiber, A. M. "Computerized Storage and Retrieval of Case Law without Indexing; the Hebrew Responsa Project." *Law and Computer Technology*. 2:14–21; 1969.
9. Bing, J. "Text Retrieval in Norway." *Program*. 15(3):150–162; 1981.
10. Bing, J.; Selmer, K. *A Decade of Computer and Law*. Oslo, Norway: Norwegian University Press; 1980.
11. Blair, D. C.; Maron, M. E. "An Evaluation of Retrieval Effectiveness for a Full-Text Document-Retrieval System." *Communications of the ACM*. 28:289–299; 1985.
12. Durkin, K.; Egeland, J.; Garson, L. R.; Terrant, S. W. "An Experiment to Study the Online User of a Full Text Primary Journal Database." In: *Proceedings of the Fourth International Online Information Meeting, London, Dec. 9–11 1980*. Oxford, England: Learned Information, Ltd.; 1980:53–56.
13. Terrant, S. W.; Garson, L. R.; Meyers, B. E. "Online Searching Full Text of American Chemical Society Primary Journals." *Journal of Chemical Information and Computer Science*. 24:230–235; 1984.
14. Franklin, J.; Buckingham, M. C.; Westwater, J. "Biomedical Journals in an Online Full Text Database; a Review of Reaction to ESPL." In: *Proceedings of the Seventh International Online Information Meeting, London, Dec. 6–8, 1983*. Oxford, England: Learned Information, Ltd.; 1983:407–410.
15. O'Connor, J. "Retrieval of Answer-Sentences and Answer-Figures from Papers by Text Searching." *Information Processing and Management*. 11:155–164; 1975.
16. O'Connor, J. "Data Retrieval by Text Searching." *Journal of Chemical Information and Computer Science*. 17:181–186; 1977.
17. O'Connor, J. "Answer-Passage Retrieval by Text Searching." *Journal of the American Society for Information Science*. 31:227–239; 1981.
18. Tenopir, C. "Retrieval Performance in a Full Text Journal Article Database." Ph.D. dissertation, University of Illinois, 1984.
19. Saracevic, T. "RELEVANCE: A Review of and a Framework for the Thinking on the Notion in Information Science." *Journal of the American Society for Information Science*. 26:321–343; 1975.
20. Sager, W. K. H.; Lockemann, P. C. "Classification of Ranking Algorithms." *International Forum On Information and Documentation*. 1(4):12–25; 1976.
21. Holsti, R. *Content Analysis for the Social Sciences and Humanities*. Reading, MA: Addison-Wesley; 1969.
22. Scott, W. A. "Reliability of Content Analysis: The Case of Nominal Scale Coding." *Public Opinion Quarterly*. 19:321–325; 1955.

ATTACHMENT #5 ATTACHED

ATTACHMENT #5

Letters to the Editor

Sir:

I wish to call your attention to the article by Jung Soon Ro in the March, 1988 issue of *JASIS* (Ro, 1988). At times, the paper is almost unintelligible. More importantly, there are several serious errors in the article. For example, on page 77 the author claims that a study at SSIE reported 30-40% higher average recall for full text searching than for searching on subject codes. This is just completely wrong. In actual fact the results reported by SSIE are 95% recall for subject indexing and 66% for text search. I really think you need to try to find better referees who are acquainted with the literature.

F. W. Lancaster
Graduate School of Library and Information Science
University of Illinois
Urbana, Illinois 61801

Ro, Jung Soon. (1988). An evaluation of the applicability of ranking algorithms to improve the effectiveness of full-text retrieval. I. On the effectiveness of full-text retrieval. *Journal of the American Society for Information Science, 39*, 73–78.

Sir:

Yesterday's mail brought the March issue of *JASIS*, whose lead article by Jung Soon Ro is nicely juxtaposed with Don Swanson's on the same subject. The former article begins with the assertion that

> It is generally accepted that information retrieval based on full texts of documents will result in higher recall and lower precision compared with retrieval using paragraphs, abstracts, or controlled vocabularies.

The latter contains the observations that

> . . . I suspect that the outcome of retrieval tests depends more strongly on the nature of the questions and the circumstances of the relevance judgments than on the characteristics of the systems under test. (page 94, col. 2) . . . consistently effective fully automatic indexing and retrieval is not possible. (Page 95, col. 2)

Having recently had occasion to conduct secondary research in this area for an information system expected to contain about three million documents, I would like to make some observations of a theoretical and practical nature. In the former case I will draw not only on my own investigations but also a study conducted for my firm by Davis B. McCarn, former Deputy Director of MEDLARS.

1. Theory

It was probably with tongue in cheek that Prof. Swanson proposed his Postulates of Impotence and Postulates of Fertility. However, in the former case he is sufficiently serious to warrant four amendments, which one might call Postulates of Hope:

1. Low recall combined with high precision may provide a sufficient number of relevant documents to answer the question, thereby obviating high recall (McCarn, 1988). (McCarn's paper seeks to develop a single measure of effectiveness combining recall and precision.)
2. "Text" is not an undifferentiated and self-explanatory term, but rather variously defined by the number of retrievable tokens predictably inherent in the type of document (Rowbottom & Willet, 1982).
3. "Indexing" is not an undifferentiated and self-explanatory term, but rather variously defined by the nature of the task, the education and experience of the indexer, the duration of and feedback to the indexing project, and the specific needs of the user population(s).
4. Interactive searching, together with query retention and analysis, can provide enough bites at the information apple to keep the searcher nourished.

It is interesting that Ro relies heavily on the Tenopir dissertation (1984), while Swanson comments approvingly on the article by Blair and Maron (1985). These two studies, whose results are diametrically opposed, have become touchstones for protagonists of full text and surrogation. It is also interesting that neither Ro nor Swanson cites a paper that sought to lay a theoretical groundwork for the debate (Svenonius, 1986). In addition to doing that, Svenonius has this to say about the Tenopir study:

> A reason given why the full-text method was able to extract unique documents from the database is that the vocabulary provided by the full text of a document is larger than that of any of its surrogates, i.e., its title, abstract, or descriptors; thus, this vocabulary expresses concepts not expressed by the surrogates, including more specific concepts. A second reason given for the performance of the full-text method in retrieving unique documents is somewhat worrisome. It would seem that on the database searched the full text used more synonyms than the controlled vocabulary. This is puzzling: What is a controlled vocabulary for? One is tempted to speculate that the controlled vocabulary used might not have been of the best sort. (Svenonius, 1986, p. 334)

One might add, "or its use in indexing may not have been adequately supervised."

On the first page of his paper, Ro states that

> Many efforts to test the effectiveness of the full-text retrieval system have been made on portions of the legal literature. . . . Most of these studies reported the superior effectiveness of full-text retrieval compared with the manual, conventional technique of index lookup on court decisions.

This is correct as far as it goes. In my experience, lawyers *always* prefer text to surrogation for legal documents, but *always* reverse their preference where evidentiary materials are concerned (correspondence, business documents, reports, etc.).

Another article not cited by any of these authors won the prize for the best *JASIS* paper of 1985 (Fugmann, 1985). In it Fugmann describes the specific conditions under which various indexing methods can be assumed to perform well and badly. Much of what he has to say echoes the theoretical explanation given by Blair and Maron, repeated in Swanson's article, as to why full text won't work on large databases.

In summary, until the players in this game can agree on a set of rules for determining the variables and evaluating test results, those of us who have to depend on their advice are not going to feel very comfortable in doing so.

2. Practice

There is a critical practical need for resolution of these questions with respect to large databases. One example is a federal database that will start being created in September 1990. (The subject matter relates to burial of an undesirable substance in such a manner as to guarantee that it will not interact with the surrounding environment for several thousand years.) It is anticipated that the database will contain as many as three million documents, averaging 5–10 pages; at present the plan is to enter them in full text. One of the uses of this database will be to substitute for exchange of documents by parties to litigation. Although there are no plans at present to evaluate the effectiveness of full-text retrieval from this database, users will include adversary counsel that may argue in court that the database fails to provide adequate recall and hence is inappropriate for its designed purpose. Should a court uphold counsel in this argument, the result could be a 2–4-year delay in legal proceedings, at a cost of 500 million to one billion dollars.

One is tempted to launch into a Jeremiad on organizations that undertake such programs without consulting information scientists. However, the point I want to make is that the debate over full text vs. surrogation has important economic ramifications. As the cost of computing comes down, and the feasibility of incorporating word processing and optical disk storage into computerized retrieval comes ever closer, the need for really serious study of the basic questions becomes ever more urgent.

John S. Jordan
John S. Jordan & Associates
Washington, D.C. 20007

Blair, D.C. & Maron, M.E. (1985). An evaluation of retrieval effectiveness for a full-text document-retrieval system. *Communications of the ACM* 28, 289–299.

Fugmann, R. (1985). The five-axiom theory of indexing and information supply. *JASIS*, 36, 116–129.

McCarn, D. (1988). Recall: Full text vs. surrogate databases.

Rowbottom, M. E. and Willett, P. (1982). The effect of subject matter on the automatic indexing of full text. *JASIS*, 33, 139–141.

Svenonius, E. (1986). Unanswered questions in the design of controlled vocabularies. *JASIS*, 37, 331–340.

Tenopir, C. (1984). *Retrieval performance in a full text journal article database*. Ph.D. dissertation, University of Illinois.

ATTACHMENT #6 ATTACHED

ATTACHMENT #6

COMPARATIVE EFFECTS OF TITLES, ABSTRACTS AND FULL TEXTS ON RELEVANCE JUDGMENTS (1)

Tefko Saracevic
Center for Documentation and Communication Research
School of Library Science
Case Western Reserve University
Cleveland, Ohio

Abstract

Twenty-two users submitted 99 questions to experimental IR systems and received 1086 documents as answers, receiving first titles, then abstracts, and finally full texts. Ability of users to recognize relevance from shorter formats in comparison to full text judgment was observed. Of 1086 answers evaluated, 843 or 78% had the same judgment on all three formats. Of 207 answers judged relevant from full text, 131 were judged so from titles and 160 from abstracts. Parallels between users' and IR systems' performance on shorter formats are drawn.

Introduction: Significance of the Concept of Relevance

A communication process can be thought of as a sequence of events resulting in the transmission of something called information from one object (source) to another (destination) (2). We may not know what information is but we can study some of its properties and effects. Analogies of such an approach abound--man doesn't know what for instance electricity is, but is quite well familiar with its behavior, properties, effects.

Systems whose function is the carrying out of a communication process are usually referred to as information systems. There are a variety of such systems, utilizing a variety of properties of information, in a variety of ways, and most importantly for a variety of purposes. Our interest here is limited to information retrieval (IR) processes and systems which are primarily concerned with semantic properties of information. An IR process can be thought of as an instrument for providing effective contact (within a given frame of reference) between the source and destination within a communication process--that is, a process which, when properly carried out, assures that the information transmitted from the source(s) to the target(s) is relevant, i.e., results in the accumulation of knowledge at the destination. In other words, relevance may be interpreted as a measure of the effectiveness of the transmission of information in an information retrieval process. Thus, it is a fundamental concept in IR processes and systems--regardless of the name it is called, the definition it is given, or the way it is treated (or ignored), relevance is fundamental to information retrieval.

Therefore, it is significant to study relevance, e.g., to study variables affecting relevance judgments by people . . . variables such as psychological factors, format of representation of information, etc. . . . since this enhances our understanding of what makes for an effective contact in an IR process. The whole effort in IR systems is basically directed toward simulation, approximation and even prediction of users' relevance judgments. Hopefully, on the basis of knowledge we gain investigating relevance we might be able (in some distant future) to practically optimize such effective contact in a formal manner --and furthermore know when we achieve an optimum in relation to sets of given constraints and variables within a given IR system.

Background of the Study

This study is concerned with and limited to the effects on human relevance judgments of variations in the format of document representations; relevance judgments by users based on titles and abstracts are compared to those on full texts, i.e., full-text judgments were taken as standard for comparison. A relatively large number of judgments, on over 1,000 titles, abstracts, and full texts, were used in this study.

The work reported here was performed within a larger project entitled Comparative Systems Laboratory, which was concerned with a variety of theoretical, experimental and control aspects related to testing and evaluation of IR systems. The whole project, (its objectives, methodology, design, experiments, analysis of results, controls, and related studies) has been fully reported elsewhere (3). Relevance judgments on various formats of documents was one of variables investigated experimentally among a set of other variables affecting performance of IR systems. This paper deals with one variable, namely relevance judgments on differing formats of output. Therefore, presentation of methodological aspects is limited to essentials for that variable. "Format of output"--as a variable--was defined as the physical representation of documents presented to a user as an answer to a question.

Previous Experimental Work

In all of the work stemming from or related to information retrieval, or more broadly, to information science, there have been no more than a dozen experimental studies directly concerned with aspects of relevance. Only three previous experiments could be found that were devoted to a problem similar to the one reported in this study: how do different document formats affect human relevance judgment?

Rath, Resnick, and Savage (4) conducted an experiment to determine which of four types of "lexical indicators" of content could be utilized best by subjects to distinguish relevant from irrelevant documents in answering 100 short questions. The lexical indicators were: titles, auto-abstracts (10% of sentences from the text selected by machine), pseudo-auto abstracts (first and last 5% of sentences from the text) and full text. Questions were based on 21 documents taken from a 92-document population; there were fifty subjects (college students) divided into groups performing the judgment. The results indicated that there were no major differences in judgment between the groups using complete text and either kind of abstracts, but the title groups scored considerably lower, i.e., determing "usefulness" of documents from titles was low. Furthermore, the group using full text expressed the highest confidence in their judgment. Though the experiment reported here differs in important respects, its results agree to some extent with trends reported by Rath.

Resnick (5) investigated the response of regular users of a Selective Dissemination of Information (SDI) system; 400 documents sent to users as a notification in response to their profiles, i.e., questions, were divided into two groups: first, where only a title was sent and the second where an abstract in addition to the title was sent. Results indicated that there were no significant differences in the ordering rate for a "hard copy" i.e., full text, of a document, when either titles or titles plus abstracts were used for notification. Furthermore, it was found that once the full text is received, the percent of full text documents judged relevant to users' interest was not significantly different between users who ordered on the basis of titles and those who ordered on the basis of titles plus abstracts. Once users received full text they judged approximately 60% of documents as being relevant. The trends in these percentages in Resnick's study are to some extent in agreement with trends reported in this study.

Rees, Schultz, et al. (6) investigated among other variables, the effects of various document representations on relevance ratings by various medically-oriented groups differing in their medical expertise. The different representations included titles, bibliographic citations, and full texts. Representations of 16 documents were judged by: 40 medical experts (M.D.'s--researchers and non-researchers), 29 medical scientists (Ph.D.'s), 25 M.D. residents, 29 medical students and 60 medical librarians; judgments were made in relation to an elaborately described diabetes research project and recorded on an 11-point relevance scale. Librarians tended to increase the scale values of the relevance ratings from titles to citations to full texts, while others tended to decrease the ratings, i.e., on the average, titles and citations were less relevant to librarians in comparison to full texts, while to others titles and citations were more relevant in comparison to full texts. Medical experts and scientists tended to judge titles and citations quite liberally, probably estimating that the full text of a document may reveal items of interest and when they have gotten to full text, their judgment was much more stringent. Just the opposite trend was found in the experiment reported in this paper, but the experimental settings, relevance scales, definitions and variables differed to such an extent that it is hard to make any comparisons.

In general then, if any comparison can be made at all, results from two of the three studies reviewed exhibited trends that are in some agreement with the trends in results of the present study.

Methodology

Nine index files containing indexes to the same 600 documents, but differing in index language and/or sources of indexing, were assembled in order to observe the effect of various variables on performance of the experimental IR system as a whole. Documents were in the field of tropical diseases. Questions were solicited and obtained from specialists (M.D.'s or Ph.D.'s) in tropical diseases working as scientists--researchers on various aspects of tropical diseases (e.g., sources, controls, nature, effects, cures) in various laboratory institutions across the U.S. These specialists were considered as a sample of real users of an IR system, by virtue of having been asked to submit questions as they arose from their current research work and interest--"real" questions that they would or did submit to a "real" IR system or library in their field. Each question was analyzed in a variety of ways (by controlled addition of synonyms, near synonyms, and/or related terms) and searched on two levels of complexity (a very broad search and a "narrow" one, i.e., as asked in the question). Each of the nine files was searched in the above manner separately. All documents for one question retrieved as answers by any search on any file were combined into one list--i.e., a union of answers was created, regardless of their source of retrieval. This union was submitted to each user for each of his questions with the request to judge independently each of the submitted documents as being either relevant (R), partially relevant (P), or not relevant (N). The following loose definitions were provided to users along other procedural instructions:

> A "relevant" (R) document is any document which on the basis of the information it conveys, is considered to be related to your question even if the information is outdated or familiar to you.
>
> A "partially relevant" (P) document is any document which, on the basis of the information it conveys, is considered only somewhat or in some part related to your question or to any part of your question.
>
> A "non-relevant" (N) document is any document which, on the basis of the information it conveys, is not at all related to your question.

Documents as answers to one given question were submitted separately in three formats of output--first a set of titles, then a set of abstracts and finally a set of full texts--and users were asked to judge each format independently of others. All three sets of formats were sent to a user at the same time, but in separate envelopes with explicit instructions to evaluate all titles first. Instructions were to complete the evaluation of titles prior to evaluating abstracts, to mail the results of evaluation back immediately and only then proceed to abstracts. The same procedure was to be followed in evaluating abstracts and then to proceed to full texts.

Hypothesis

Specifically, the basic purpose in testing the format of output as a variable was to observe the changes in relevance judgments of users when the following representation of the output of searches was presented to the users in exact succession:

1. Titles, together with bibliographic citation (author, journal date, etc.). Bibl. cit. was used in addition to text of titles because this is a practice in "real" IR systems to which users are accustomed--we wanted to retain reality as much as possible; titles were 5 to 9 words in length.

2. Abstracts including titles with bibliographic citations. Abstracts used were photocopies from Tropical Diseases Bulletin, the abstracting journal for articles in tropical diseases from London (England). Abstracts were 250-400 words in length.

3. Full texts of documents. These were photocopied from journals where they appeared; they were 2000 to 4000 words in length.

A null hypothesis was stated: different formats of output do not affect the relevance judgments of users.

The three formats were chosen so that it might be possible to observe the relationship between user evaluation and document format, where the format was varied to allow for a gradual increase in length, assumed to be a gradual increase in the information presented. The full text provides the maximum length and maximum information available; thus the judgment of the full text was taken as final. If so, then the user's judgment of the shorter representation, titles, and abstracts, can be expected to differ occasionally; in a sense, to "err" in two directions: leniency or strictness. Such "errors" seem to indicate an inability on the part of the user to determine final relevance from the shorter representations. This should not, however, be construed as a test of the user. Naturally, the user's ability is largely dependent upon the degree to which shorter formats accurately represent the content of the full text. In a sense, we may consider this as a user's performance on shorter formats.

The above reasoning rests on three additional assumptions: first, that a user does make the judgment on the text content of a shorter representation and not occasionally on some other clues, such as author, journal, date; second, that the judgment on each document in the set is entirely independent from judgments on other documents in the set; and third, that the manner of presentation of the sets of documents of differing formats made it possible for the judgments on one format to be independent of other formats, i.e., that the users followed the instructions fully.

Method of Comparison and Analysis

The first method of comparison of three judgments that came to mind was to add up the judgments (i.e., number of answers judged relevant (R), partially relevant (P), and not relevant (N)) for each format separately over all questions. The sum of judgments on shorter formats could then be compared with the sum of judgments on full texts; e.g., the number of answers judged R on titles could be compared with the number of answers judged R on full texts. However, this method does not provide a complete picture of the ability of the users to determine relevance from shorter formats because it will not indicate whether the membership of the set of answers judged relevant varies from format to format. For example, of a total of eight answers submitted, a user may judge four answers relevant and four non-relevant on all formats, but the four relevant answers in his final judgment on full text may have been judged not relevant in a previous judgment on a shorter format, or vice versa.

To obtain a more accurate picture of the situation, it is necessary to tabulate for each pair of formats the intersections and changes of all possible relevance judgments. Since a three-point scale for judging relevance (R, P, and N), was used, there are nine possible intersections and changes of judgments between any two formats. The 3 x 3 table in Figure 1 shows the possible intersections and changes:

Figure 1

		Judgments on Full Texts (Final Judgment)		
		R	P	N
Judgments on Shorter	R	R-R	R-P	R-N
Representation	P	P-R	P-P	P-N
(or 1st Judgments)	N	N-R	N-P	N-N

This is, of course, also a general table that can be used on any pair of relevance judgments. As used here, indicated along the side of the table (when numbers will be substituted for letters), will be the number of answers judged R, P, N on a shorter format (e.g., titles) and along the top of the table will be the number of answers judged R, P, N on full texts. The entries in the cells represent the number of answers judged R (or P or N) on the shorter format that were judged R (or P or N) on the full texts. For example: in the upper left corner the entry, R-R will mean the number of answers judged relevant on titles (or abstracts) which were also judged relevant on full texts; in the lower left corner the entry, N-R, will mean the number of answers judged not relevant on titles (or abstracts) but judged relevant on full texts. The diagonal, R-R, P-P, N-N, represents instances where the user did not change his judgment. All other cells represent changes.

Measures of User's Performance

In order to be fully exploited for conclusions, data, as displayed in tables of form in Figure 1, should further be expressed in terms of some probabilities, ratios, percentages, or proportions, etc., that is in terms of some measures. No specific measures expressing users' performance were readily available. Therefore, it was decided to use the same measures that were used to express effectiveness-related performance of IR systems, but proper operational redefinitions were necessary. This approach, using the same measures indicating systems' as well as users' performance, led to an examination of interesting parallels and to subsequent interesting conclusions about performances of IR systems.

Measures of effectiveness of IR systems performance used in connection with the Comparative Systems Laboratory were Sensitivity (Se), Specificity (Sp) and Effectiveness (Es). (7)

Sensitivity was defined as the conditional probability that a member (document) of a file will be retrieved by a system if it is relevant. Operationally, it can be approximated as the ratio between relevant documents retrieved and all relevant documents in the file. Specificity was defined as the conditional probability that a member (document) of a file will not be retrieved (i.e., will be suppressed) if it is not relevant. Operationally, it can be approximated as the ratio between non-relevant documents not retrieved and all non-relevant documents in the file. Effectiveness is a function of Se and Sp such that: Es = Se + Sp - 1.

Since we have chosen to express the user's ability to determine relevance of a document from differing representations by the same measures, what remains to be done is to interpret specifically the meaning of the measures in the context of this experiment. As mentioned, we called the judgment on full text the final relevance evaluation of the answer. Thus the value of Se, Sp or Es as a measure of the users' ability to determine the relevance of a document from full text representation is inevitably 1.00. Se, Sp and Es for shorter representations mean as follows:

1. Sensitivity (Se) measures the users' ability to recognize relevant answers by their titles or abstracts in comparison to full text judgments. Se of 1.00 means that all documents judged relevant on full texts were also judged relevant on titles or abstracts, and Se of 0.00 means that none were judged relevant. Operationally, Se was approximated as the ratio between the number of documents judged R on a shorter format that were also judged R on full text and the number of documents judged R on full text.

2. Specificity (Sp) measures the users' ability to recognize non-relevant answers by their titles or abstracts. Sp of 1.00 means that all documents judged non-relevant on full texts were also judged non-relevant on titles or abstracts, and Sp of 0.00 means that none were judged non-relevant. Operationally, Sp was approximated as the ratio between the number of documents judged N on shorter format that were also judged N on full text and the number of documents judged N on full text.

3. Effectiveness (Es) is a function of Se and Sp such that: Es = Se + Sp - 1.

Prior to actual calculations, however, an additional problem, reflecting a weakness of the operational definition of measures had to be solved: the treatment of partially relevant documents (P's). The measures as operationally defined can be calculated only from a two-point relevance scale (R and N) and not from a three-point relevance scale (R, P, and N), as used in the experiment and as shown in Figure 1. Thus, the partially relevant judgments must be reinterpreted in order to arrive at a two-point scale. Specifically, P's can be added either to R's, or to N's, or excluded altogether. To accomodate all three possibilities, three sets of measures were calculated:

1. Se_1, Sp_1, and Es_1, where the partially relevant judgments were added to non-relevant ones (P added to N)

2. Se_2, Sp_2, and Es_2, where the partially relevant judgments were added to relevant ones (P added to R)

3. Se_3, Sp_3, Es_3, where the partially relevant answers were excluded entirely and calculations done only on the basis of actual R's and N's.

Now let us draw parallels between the users' performance and an IR system performance. If a user is able to determine his final evaluation from a shorter format, i.e., if his judgment did not change from a shorter format to a full text (final) judgment, then a given system could reasonably be asked to attempt to do the same with indexes based on that same shorter format. If, however, the user was not able to do this, i.e., if his judgment changed, then in general, a system cannot reasonably be expected to do better than a user in assigning relevance to documents from indexes based on that format. Thus, the users' performance on shorter formats can be considered a reasonable goal for a system performance on shorter formats. Conclusions to that effect will be made later in the paper.

Results

The data base for comparison was arrived at by adding up the total number of documents and the R, P, N judgments on those documents submitted as answers over all the questions. There were 22 users who submitted 99 questions; altogether for those 99 questions there were a total of 1086 answers judged by users on all three formats (i.e., for one question there may be 30 answers, for another 17 . . . and so forth . . . when all added up over 99 questions the number of answers was 1086).

The judgments were distributed over the three formats as follows:

	No. of Answers Judged Relevant	Partially Relevant	Not Relevant	Total
Titles	167	157	762	1086
Abstracts	175	169	742	1086
Full Text	207	156	723	1086

Figure 2 presents the intersections and changes of relevance judgments between titles and full texts. The arrangement of this table was explained in full while treating Figure 1. An example from reading the table: the first row indicates that of the 167 answers judged R from titles, there were 131 judged R, 13 P and 23 judged N from full texts. The first column indicates that of the 207 answers judged R from full text there were 131 judged R, 33 P, and 43 judged N from titles:

Figure 2

		Full Text Judgments		
		207 R	156 P	723 N
Title Judgments	167 R	131	13	23
	157 P	33	95	29
	762 N	43	48	671

Figure 3 presents the intersections and changes of relevance judgments between abstracts and full texts. The arrangement is the same as in the previous figure:

Figure 3

		Full Text Judgments		
		207 R	156 P	723 N
Abstract Judgments	175 R	160	3	12
	169 P	23	125	21
	742 N	24	28	690

Figure 4 presents the measures of $Se_{1,2,3}$, $Sp_{1,2,3}$, and $Es_{1,2,3}$, indicating the users' performance on judgments between titles and full texts, where the full texts judgments are taken as final and thus as a standard. The measures, as explained, indicate the users' ability to approximate the same judgment from titles as had been done from full texts:

Figure 4

Se_1 Sp_1 Es_1	Se_2 Sp_2 Es_2	Se_3 Sp_3 Es_3
.63 .96 .59	.75 .93 .68	.75 .97 .72

Figure 5 presents the same measures but indicates the users' performance on judgments between abstracts and full texts, where the full text judgments are taken as final and thus as a standard:

Figure 5

Se_1 Sp_1 Es_1	Se_2 Sp_2 Es_2	Se_3 Sp_3 Es_3
.77 .98 .75	.86 .95 .81	.87 .98 .85

No statistical tests for acceptance/rejection of null hypothesis were used for two reasons: first, it is felt that the data does not conform to the assumption under which the most interesting tests, such as Friedman analysis of variance, are run (primarily the data is not independent) and second, interesting conclusions can be made from a direct inspection of data without recourse to statistical tests of significance.

Conclusions

1. Null hypothesis: From the direct inspection of the data it may be decided even without formal statistical tests to reject the null hypothesis, which states that the formats of output do not effect the relevance judgments of the user. It seems that different representations of documents significantly affect the users' relevance judgment. However, the set of answers judged "relevant" in various formats differed from format to format more than the set of answers judged "not relevant." It seems to be easier for the users to recognize non-relevance from the shorter formats than relevance; all this, of course, with the judgment of full text taken as final and as the standard to which other judgments are compared.

   Thus, if the null hypothesis had been worded in a more restrictive form, such as: "The formats of output do not effect the non-relevant judgments of the user," it could not have been clearly rejected, nor clearly accepted.

2. Immutability: The overall percentage of immutability from titles to full text was 85%, which means that 897 of the 1086 documents submitted as answers had the same judgment on titles as on full text. The immutability from abstracts to full text was 90%, i.e., 975 of the 1086 answers had the same judgment. The immutability of judgment from titles to abstracts to full text was 78%, 843 of the 1086 answers had the same judgments on all three formats. Thus, the answers that had different judgments on different formats constitute 22% of the total, or 243 answers. This can be considered a significant number, since over 1/5 of the output had different judgments on different formats.

In general, the shorter the representation in comparison to full text, the more changes in judgments can be expected.

3. Sensitivity: All three Se measures for titles are significantly smaller than those for abstracts. The highest Se for both titles and abstracts is Se_3, which entirely excludes the partially relevant judgments from the calculations. The lowest Se for both representations is Se_1, which includes P's with N's. In that lowest case, the Se_1 of .63 for titles is low indeed, meaning that only 131 answers (i.e., 63%) were judged clearly relevant (R) on titles of the 207 judged so on full text. For abstracts Se_1 is somewhat better, being .77, or 160 of the 207 answers were judged clearly relevant on both abstracts and full text.

4. Specificity: The Sp values for titles are somewhat smaller than those for abstracts. The highest Sp is Sp_3 which, as the highest Se, excludes P's entirely. The lowest Sp is Sp_2, where the P's are counted with R's. The lowest Sp for titles was .93, or 671 answers were judged not relevant from titles of the 723 judged so from full texts. Sp_2 for abstracts was .95, or 690 of the 723 answers were judged not relevant on abstracts and full texts. In other words, the differences between non-relevant judgments on titles and abstracts were not great. The high number of non-relevant answers submitted to users stems primarily from broad searches which were designed basically as a "pulling out" device for all existing potentially relevant answers in the files.

5. Effectiveness: Abstracts were considerably more effective than titles. The highest Es of .85 for abstracts and .72 for titles was Es_3, where P's were excluded, and the lowest was Es_1 (.76 for abstracts and .59 for titles), where P's were treated as N's.

In general, if given a choice, the judgment from abstracts should clearly be preferred over the judgment from titles.

6. Partially Relevant Judgments: The exclusion of P's in calculations increased all three measures used. This indicates that P's have a special, unstable property, wandering the most widely over all of the judgments.

Discussion

A few positive aspects clearly stand out: relevance judgments seem to agree to a considerable extent over various formats, but not completely. The degree of agreement encourages the notion of a certain stability and reliability of relevance judgments, and thus this may justify their applicability as more or less objective criteria for measures. It seems, as concluded from a number of relevance experiments, that people under well-defined conditions and within acceptable limits can be objective measuring instruments indicating relevance, objective because a relatively high correlation (as human agreement correlations go) of agreement exists.

Turning to generalizations from the results: if an IR system requirements can be satisfied by the user recognizing approximately 2/3 to 3/4 of relevant answers of all possible relevant answers, then the titles can be submitted as output format as answers to questions. This conclusion is, of course, of special interest to KWIC or other systems which distribute titles only. The abstract will provide approximately 4/5 to 5/6 recognition in comparison to the 100% recognition from full text. Clearly, the cost of submitting titles is much lower than that of submitting abstracts which, in turn, is considerably lower than that of submitting full texts. So taking cost into account, submitting titles or abstracts is not a "bad deal," especially if knowing that most of the answers submitted will be non-relevant anyway. The cost of throwing away non-relevant titles or abstracts is much less than that of throwing away full texts.

Now to turn to the problem of partially relevant answers. They displayed a high degree of instability over judgments and, as witnessed by the performance of other variables in the experiments not treated here (3), having special characteristics, they were not easy to handle. This brings us to the following generalization: it seems that there exists among the answers supplied by a system to a user a clear set of "hard core" relevant answers, a clear and large set of "hard core" non-relevant answers, and a fuzzy gray area of answers that are neither clearly relevant nor clearly non-relevant. The answers from the last set can be considered as having various "degrees" of relevance, while the other two sets, clearly relevant and clearly non-relevant, can be thought of as the opposite ends of a continuous relevance scale.

Finally, let us compare user and system performance on different document formats. It was mentioned that if a user cannot determine relevance from a shorter format in relation to final relevance of full text, a system cannot be expected to do any better. The Comparative Systems Laboratory investigated, among other variables, the performance of index files based on titles, abstracts, and full text. For simplicity, let us compare only the measures where P's were treated as N's. Se_1 of users on titles was Se_1 - .63 and the highest Se_1 of title-based index files in CSL was .37, achieved by the broad search. Se_1 of users on abstracts was .77 and the highest Se_1 of abstract-based index files was .80.

Thus, as far as the recognition/retrieval of relevant answers is concerned, the users performed considerably better on titles than systems, and both did almost the same on abstracts. The logical explanation for the considerably higher performance by users on titles is that the users "read into or from" the titles (which included bibliographic citations) additional information based on their own knowledge of the field, enabling them to estimate relevance of a document better than IR systems based on the same format. In the case of abstracts, it seems that both users and systems utilized the same information and the system, as designed, performed with respect to Se as closely as can be expected.

Regarding specificity, the users performed in all cases considerably better than the experimental IR systems used in CSL: i.e., they were much better in recognition of non-relevant answers than systems were. Therefore, users also achieved a higher effectiveness on performance than did systems.

In general, then, a title based index file cannot be expected to perform any higher than .60 - .75 sensitivity and abstract-based index files any higher than .75 - .85 sensitivity--because the users do not recognize relevant answers from comparable formats at any higher rate. These generalizations assume, of course that the users' judgments of full texts is final and standard, which in any case seems a reasonable assumption.

References

1. Work performed under PHS Grant FR-00118. Without the work on organization of data by Carolyn Gifford and George J. Baumanis this paper would not be possible. Their contribution is gratefully acknowledged.

2. Goffman, W., and V. A. Newill., "Communication and Epidemic Processes." Proceedings of the Royal Society, Series A. Vol. 298, No. 1454, 2 May 1967, pp. 316-334.

3. [redacted]An Inquiry into Testi of Information Retrieval Systems. 3 Parts. Comparative Systems Laboratory Final Report. Center for Documentation and Communication Research, Case Western Reserve [redacted] Cleveland, Ohio, 1968, 611 p. [redacted] PB [redacted]

4. Rath, G. J., A. Resnick, and T. R. Savage, "C[redacted]pes of Lexical Indicators." American Documentation, Vol. 12, No. 2, April 1961, pp. 126-130.

5. Resnick, A., "[redacted] of Document Titles and Abstracts for Determining Relevance of Documents." Science, Vol. 134, No. 3484, 6 Oct. 1961, pp. 1004-1006.

6. Rees, A. M., D. G. Schultz, et al., A Field Experimental Approach to the Study of Relevance Assessments in Relation to Document Searching. 2 vols. Center for Documentation and Communication Research, Case Western Reserve University, Cleveland, Ohio, 1967, 475 p. [redacted].

7. Goffman, W. and V. A. Newill, "Methodology for Test and Evaluation of Information Retrieval Systems." Information Storage and Retrieval, Vol. 3, [redacted].

asis

AMERICAN SOCIETY FOR
INFORMATION SCIENCE

PROCEEDINGS of the AMERICAN SOCIETY FOR INFORMATION SCIENCE

Volume 6

COOPERATING INFORMATION SOCIETIES

Edited by JEANNE B. NORTH

32ND ANNUAL MEETING
SAN FRANCISCO, CALIFORNIA
OCTOBER 1-4, 1969

GREENWOOD PUBLISHING CORPORATION
Westport, Connecticut • London
GP

ATTACHMENT #7 ATTACHED

8-28-90

ATTACHMENT #7

LICENSING SUPPORT SYSTEM
REVISED DATA SCOPE ANALYSIS

Draft

SCIENCE APPLICATIONS INTERNATIONAL CORPORATION

August 1990

1.0 INTRODUCTION

The purpose of this analysis is to review and update the information presented in the Licensing Support System Preliminary Data Scope Analysis (DOE, 1989a) to determine if significant changes have occurred in the projected number of pages to be loaded into the Licensing Support System (LSS). The analysis was performed in two steps. First the major sources of data were examined to determine the current (1990) information believed to be relevant to the LSS database and the rate at which that information is being accumulated. Second, a projected rate at which pages would be generated for LSS was developed from the current official schedule for the DOE OCRWM high level waste repository program. These factors were used to generate a new table of pages to be loaded into the LSS through the year 2009.

1.1 Background

The Licensing Support System Preliminary Data Scope Analysis was developed in early 1988 and provided an estimate of the LSS data base in August 1990 and projections through 2009 based on the following sources:

1) NNWSI Project Participants and Subcontractors

2) NNWSI ARS data base

3) OCRWM Headquarters ARS data base

4) NRC data base

5) Regulations

6) Commitments Tracking data base

In addition to these sources, an adjustment was added to account for an estimated under-representation of relevant topics. The final results, as presented as Table 8 of the LSS Preliminary Data Scope Analysis, indicated a cumulative page count at the end of 2009 of between 30,947,000 and 40,567,000. Subsequent to the completion of that report an error was discovered in the analysis and a revised Table 8 was produced in the Licensing Support System Conceptual Design Analysis (DOE, 1989b) which shows an increase in the 2009 page count to between 32,191,000 and 42,216,000.

Upon further review of these figures, an additional mathematical error was discovered in the calculations of the projected data scope size in August 1990 as defined in the LSS Preliminary Data Scope Analysis. Item 7 of Table 7 (page 42) calculated the contribution of the ARS/Washington documents to be 323,000 pages from March 1988 to August 1990, but the correct figure is 32,300 pages. Since the adjustment for under-estimation of topics (Item 13 of Table 7) was calculated as a percentage of the contributions from the various sources, it

Draft

must also be corrected. This results in the 1988 estimate of pages in August 1990 to be lowered slightly to 8,705,000 pages for the low estimate and 10,705,000 pages for the high estimate. These corrected figures were then used to recalculate the projections for 1990 through 2009. The results are shown in Table 1-1, 1988 Projection of the Size of the LSS Data Base. Table 1-1 is then the correct basis for comparison with the current analysis.

1.2 Analysis of Contributions

It is helpful to review the projected sources of information for the LSS data base and their relative size. Based on the information from the LSS Preliminary Data Scope Analysis (DOE, 1989a) as corrected above, the major data sources for the size of the August 1990 data base and their relative contributions are summarized below:

1) NNWSI Project Participants and Subcontractors

Estimated 1980 to August 1990 production at 4,233,000 pages, all of which was considered relevant to LSS.

2) NNWSI ARS Data Base

Existing (1988) data base plus production to August 1990 estimated at 86,700 documents. In addition a backlog of 845,000 documents had not yet been entered into the system. At 8 to 10 pages per document, a judgement of relevance at 65% to 70%, and 60% considered to be non-duplicative of the NNWSI participants contribution, the total ranged from 2,906,000 to 3,913,000 pages.

3) OCRWM Headquarters Data Base

Existing (1988) data base of 113,000 documents is supplemented by the estimated production to August 1990 of 29,000 additional documents and a backlog of 162,000 documents that had not yet been loaded into the system. At 5.8 pages per document, an estimated 20% relevance, and only 4% considered to be duplicated in the above contributions, the total Headquarters contribution is estimated at 338,000 pages.

4) NRC Data Base

The existing data base plus backlog in 1988 was estimated at 50,000 documents plus an additional 17,400 document estimated to be generated between 1938 and August 1990. At 7 pages per document and an estimated 90% relevance, the total contribution to the August 1990 data base is 425,000 pages.

TABLE 1-1. 1986 PROJECTION OF THE SIZE OF THE LSS DATA BASE

	LOW ESTIMATE		HIGH ESTIMATE	
Year	Pages Added During Year	Cumulative Pages At Year-End	Pages Added During Year	Cumulative Pages At Year-End
1990	830,000	8,982,000	1,100,000	11,533,000
1991	1,087,000	10,069,000	1,441,000	12,974,000
1992	1,428,000	11,497,000	1,892,000	14,866,000
1993	1,660,000	13,157,000	2,200,000	17,066,000
1994	2,009,000	15,166,000	2,662,000	19,728,000
1995	1,858,000	17,024,000	2,463,000	22,191,000
1996	1,635,000	18,659,000	2,167,000	24,358,000
1997	1,386,000	20,045,000	1,837,000	26,195,000
1998	1,037,000	21,082,000	1,374,000	27,569,000
1999	1,286,000	22,368,000	1,704,000	29,273,000
2000	1,170,000	23,538,000	1,550,000	30,823,000
2001	1,877,000	25,415,000	2,487,000	33,310,000
2002	1,236,000	26,651,000	1,638,000	34,948,000
2003	1,261,000	27,912,000	1,671,000	36,619,000
2004	1,327,000	29,239,000	1,759,000	38,378,000
2005	1,120,000	30,359,000	1,484,000	39,862,000
2006	415,000	30,774,000	550,000	40,412,000
2007	365,000	31,139,000	484,000	40,896,000
2008	365,000	31,504,000	484,000	41,380,000
2009	365,000	31,869,000	484,000	41,864,000

Draft

These data sources are summarized in Table 1-2, along with the minor contributions from the regulations and commitment tracking, and the relative percentage of these sources are shown.

TABLE 1-2. CONTRIBUTIONS TO THE AUGUST 1990 DATA BASE PROJECTION

	Low Estimate		High Estimate	
	Pages	Percent	Pages	Percent
NNWSI Project	4,233,000	49	4,233,000	40
NNWSI ARS	2,907,000	33	3,913,000	36
Headquarters ARS	338,000	4	338,000	3
NRC	425,000	5	425,000	4
Misc	11,000	0	12,000	0
Adjustment	791,000	9	1,784,000	17
Totals	8,705,000		10,705,000	

From the above it can be seen that the contributions from the Yucca Mountain project (Las Vegas) make up approximately 80 percent of the total. Therefore the accuracy of the figures from this source are more important than the accuracy of contributions from OCRWM headquarters, for example. The miscellaneous contributions from regulations and commitment tracking are smaller than the uncertainties in other contributions and can be ignored. (Commitment tracking has been eliminated from the LSS scope).

2.0 DATA BASE ESTIMATE, 1990

The first step in the analysis is to determine the size of the data base that could be loaded into the LSS in 1990 and the rate at which that data base is growing. In order to evaluate this, the content of the major data bases that encompass information that is a candidate for inclusion into the LSS is examined, along with the associated growth rates, and an estimate of relevance and non-duplication is applied. No adjustment is applied for underestimation of topics as was done in the previous analysis. The topics to be included in the LSS were defined as a part of the negotiated rule-making process, and these definitions were used to assess the relevance factors that are applied in this analysis.

2.1 Yucca Mountain Project

Since the 1988 analysis, the Yucca Mountain Project Office has accumulated the products of the participating subcontractors and has entered, or is in the process of entering, records of these documents into the Records Information System (RIS). Therefore it is not necessary to separate the Project Office source from the participating subcontractors. As of March 1990, the RIS contained approximately 100,000 records representing 1,725,000 pages (based on microfilm frame count). In addition to the records in the system, an additional 5,500,000 pages of information has been accumulated but not yet entered. New information is being accumulated at an average rate of 1500 records per week. Based on an average 17 pages per record (or document), the rate of new pages generated is 1,326,000 pages per year.

Much of the technical information in the Yucca Mountain data base was performed before the Quality Assurance Program was operative. Therefore it is difficult to ascertain whether or not this information will be entered into the LSS. In keeping with the philosophy of the 1988 analysis, the page count will be estimated as both a "low estimate" and a "high estimate". For the low estimate it is assumed that this technical data will not be entered into the LSS. With that assumption, it is estimated that only 10 percent of the project data base is relevant. For the high estimate, assuming the technical data is entered into the LSS, the relevance of this data is estimated at 75 percent.

For most of 1990, no significant technical analysis has taken place. Most of the generation of paper pertains to the development of strategies in the legal arena. Therefore it will be assumed that for both the low and high estimates, the percent relevance for 1990 is 10 percent.

In 1991 site work is expected to begin. When this work is in full operation, the information generated in the project is assumed to be 75 percent relevant to LSS. Therefore, from 1992 onward, the low and high estimates will be based on the assumption that 75 percent of the documents entering the Las Vegas RIS will be entered into LSS. For 1991, a transition figure of 45 percent relevance is assumed.

2.2 DOE Headquarters

As of April of 1990, the DOE Headquarters RIS contained 192,402 records. On the basis of the number of microfilm rolls corresponding to those records, and using 3000 frames per roll, those records represent 1,205,000 pages or approximately 6.3 pages per record. Based on the first four months of 1990, the number of records processed per week (representing new document generation) are 358 corresponding to 5431 pages. (Note that this average number of pages per record for this period at 15.2 is significantly higher than the cumulative average to date of 6.3.) The annual rate of pages entered into the Headquarters RIS is then 282,000.

With respect to the applicability of the headquarters records to the LSS, the estimate is that only 15 percent of the RIS information would be relevant for material collected to date and until site work starts in 1991. After that time, there will be periods of time when the M&O Contractor will generate significant and relevant documents. Therefore, the relevance of documents will be estimated at 20 percent from 1992 onward for the low estimate, and 35 percent for the high estimate. For 1991 the figures will be 15 percent and 25 percent respectively. In addition, a factor of 0.9 will be applied to all pages from this source to account for an estimated 10 percent duplication of records with the Yucca Mountain Project contribution.

2.3 Nuclear Regulatory Commission

Estimates of documents and pages within the Nuclear Regulatory Commission data base (NEWDOCS) were received, calculating only those documents which are assumed to be non-duplicative and relevant to LSS. These figures are 157,500 pages currently cataloged and a rate of accumulation of 20,250 pages per year. These figures are significantly below the figures used in the previous analysis.

2.4 Total Contributions

Normalizing all data to the end of 1990, the summary of the contributions from the three major sources appears in Table 2-1. At the end of 1990, these estimates indicate that the candidate pages for the LSS range from 1,173,000 pages to 5,654,000 pages. The comparative figures from the 1988 analysis are from 8,982,000 pages to 11,533,000 pages.

Table 2-1. SUMMARY OF LSS DATA SOURCES

A. LOW ESTIMATE

	Yucca Mount.	DOE HQ	NRC
Pages, end 1989	6,894,000	1,118,000	142,000
Relevance, pre 1990	0.10	0.15	1.0
Non-duplicative	1.0	0.90	1.0
LSS Pages, end 1989	689,000	151,000	142,000
Pages, 1990	1,326,000	282,000	20,000
Relevance, 1990	0.10	0.15	1.0
Non-duplicative	1.0	0.90	1.0
LSS Pages, 1990	133,000	38,000	20,000
LSS Contribution, end 1990	822,000	189,000	162,000

B. HIGH ESTIMATE

	Yucca Mount.	DOE HQ	NRC
Pages, end 1989	6,894,000	1,118,000	142,000
Relevance, pre 1990	0.75	0.15	1.0
Non-duplicative	1.0	0.90	1.0
LSS Pages, end 1989	5,171,000	151,000	142,000
Pages, 1990	1,326,000	282,000	20,000
Relevance, 1990	0.10	0.15	1.0
Non-duplicative	1.0	0.90	1.0
LSS Pages, 1990	133,000	38,000	20,000
LSS Contribution, end 1990	5,304,000	189,000	162,000

3.0 DATA BASE GROWTH THROUGH 2009

Two factors affect the growth of the data base from 1990. As mentioned in the previous section, the relevance of the Yucca Mountain contribution will grow to 45 percent in 1991 and to 75 percent from 1992 through 2009. Similarly the relevance of the DOE Headquarters contributions will change with time in 1991 and 1992. More importantly there are phases in the project schedule when significant activities occur that result in increased generation of data that is candidate for inclusion into the LSS. Figure 3-1 illustrates that relative project activity and relates the activity to milestones in the project schedule as defined by DOE in November, 1989 (DOE, 1989c). This activity base does not include any milestones or activities relative to the Monitored Retrieval Storage (MRS).

From the project activity projection, the relative generation of candidate documents was calculated compared to the generation rate in 1990. Applying these project activity factor and the relevance factors results in a new estimate of pages for the LSS data base as a function of time. The results as shown in Table 3-1 may be compared with the previous results illustrated in Table 1-1. Figure 3-2 is a graphical plot of the cumulative pages at the end of each year for both the old and new estimates.

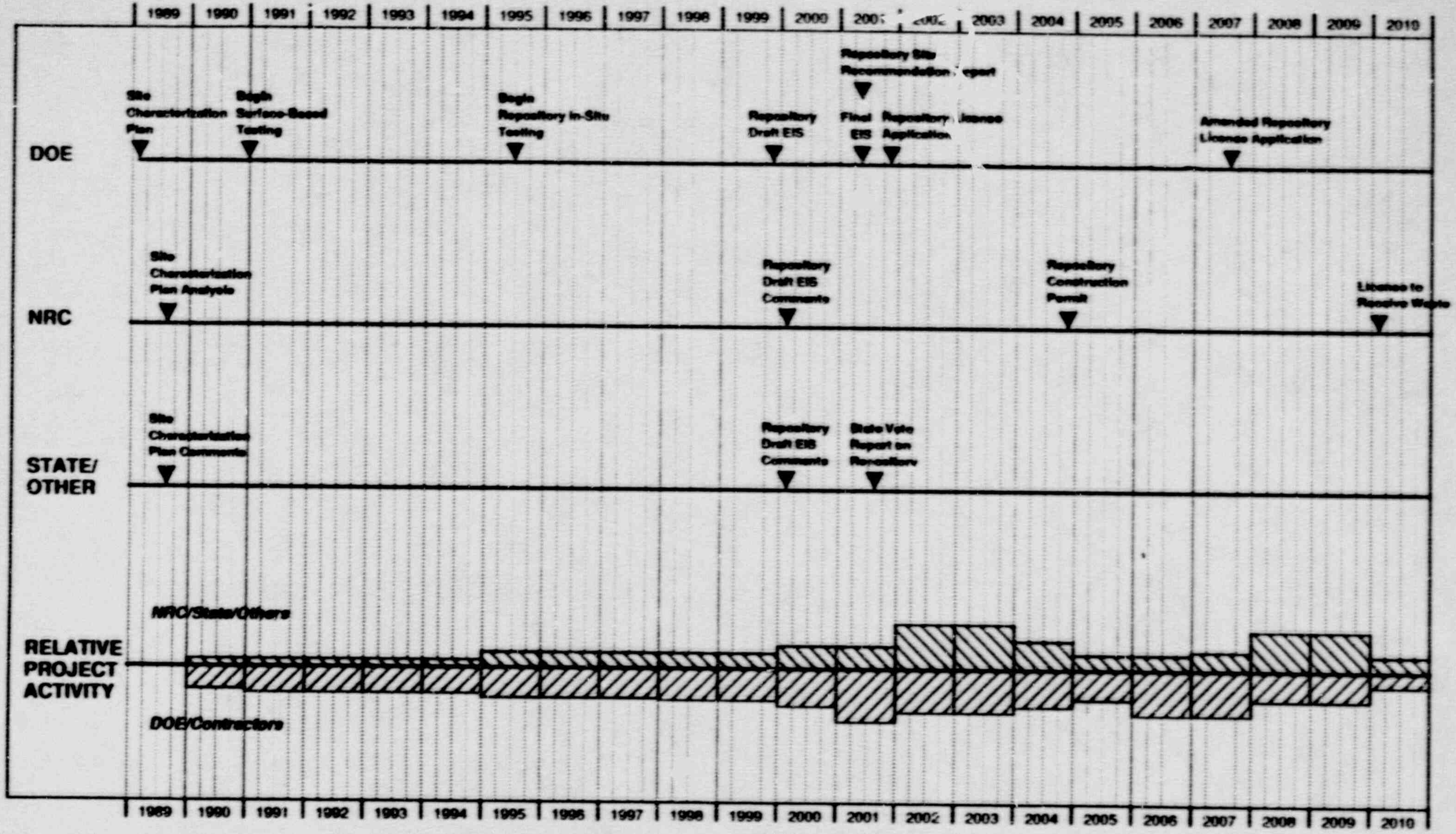


Figure 3-1. Formal Deliverables Impacting LSS

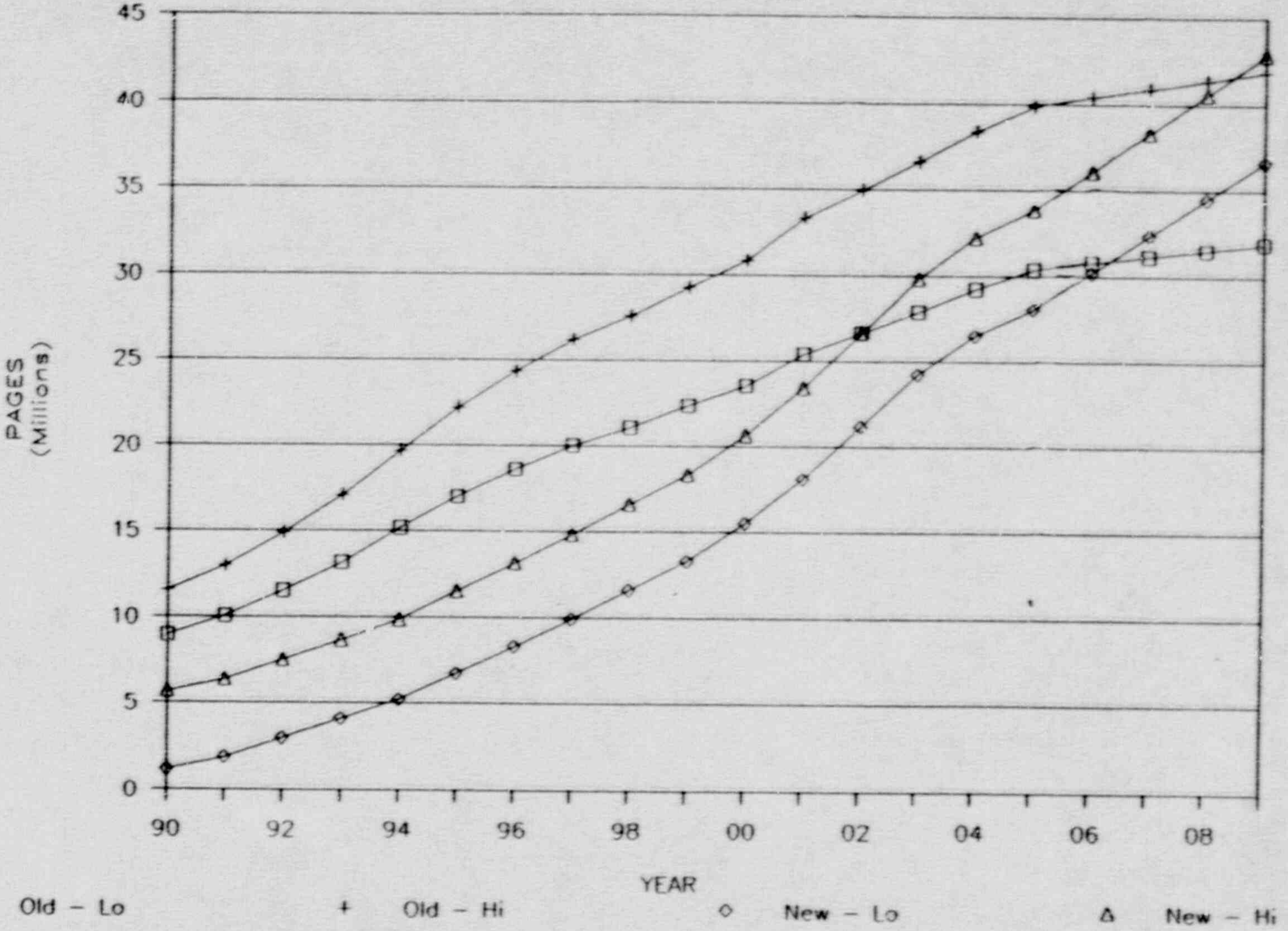
TABLE 3-1. 1990 PROJECTION OF THE SIZE OF THE LSS DATA BASE

	LOW ESTIMATE		HIGH ESTIMATE	
Year	Pages Added During Year	Cumulative Pages At Year-End	Pages Added During Year	Cumulative Pages At Year-End
1990	191,000	1,173,000	191,000	5,655,000
1991	688,000	1,861,000	715,000	6,370,000
1992	1,106,000	2,967,000	1,159,000	7,529,000
1993	1,106,000	4,073,000	1,159,000	8,688,000
1994	1,106,000	5,178,000	1,159,000	9,847,000
1995	1,580,000	6,758,000	1,656,000	11,503,000
1996	1,580,000	8,338,000	1,656,000	13,158,000
1997	1,580,000	9,917,000	1,656,000	14,814,000
1998	1,685,000	11,682,000	1,766,000	16,580,000
1999	1,685,000	13,287,000	1,766,000	18,346,000
2000	2,211,000	15,498,000	2,318,000	20,664,000
2001	2,633,000	18,131,000	2,760,000	23,424,000
2002	3,054,000	21,185,000	3,201,000	26,625,000
2003	3,054,000	24,239,000	3,201,000	29,826,000
2004	2,243,000	26,482,000	2,351,000	32,177,000
2005	1,580,000	28,061,000	1,656,000	33,833,000
2006	2,106,000	30,168,000	2,208,000	36,041,000
2007	2,159,000	32,327,000	2,263,000	38,304,000
2008	2,159,000	34,485,000	2,263,000	40,567,000
2009	2,159,000	36,644,000	2,263,000	42,829,000

Figure 3-2. COMPARISON OF OLD AND NEW DATA SCOPE ANALYSES

Figure 3-2. COMPARISON OF OLD AND NEW

DATA SCOPE ANALYSES



REFERENCES

DOE, 1989a; Licensing Support System Preliminary Data Scope Analysis, U.S. Department of Energy, Office of Civilian Radioactive Waste Management, DOE/RW-XXXX, January 1989

DOE, 1989b; Licensing Support System Conceptual Design Analysis, U.S. Department of Energy, Office of Civilian Radioactive Waste Management, DOE/RW-XXXX, January 1989

DOE, 1989c; Report to Congress on Reassessment of the Civilian Radioactive Waste Management Program, U.S. Department of Energy, Office of the Civilian Radioactive Waste Management, DOE/RW-0247, November 1989

ATTACHMENT #8 ATTACHED

ATTACHMENT #8

Indexing by Latent Semantic Analysis

Scott Deerwester
Center for Information and Language Studies, University of Chicago, Chicago, IL 60637

Susan T. Dumais*, George W. Furnas, and Thomas K. Landauer
Bell Communications Research, 445 South St., Morristown, NJ 07960

Richard Harshman
University of Western Ontario, London, Ontario Canada

A new method for automatic indexing and retrieval is described. The approach is to take advantage of implicit higher-order structure in the association of terms with documents ("semantic structure") in order to improve the detection of relevant documents on the basis of terms found in queries. The particular technique used is singular-value decomposition, in which a large term by document matrix is decomposed into a set of ca. 100 orthogonal factors from which the original matrix can be approximated by linear combination. Documents are represented by ca. 100 item vectors of factor weights. Queries are represented as pseudo-document vectors formed from weighted combinations of terms, and documents with supra-threshold cosine values are returned. Initial tests find this completely automatic method for retrieval to be promising.

Introduction

We describe here a new approach to automatic indexing and retrieval. It is designed to overcome a fundamental problem that plagues existing retrieval techniques that try to match words of queries with words of documents. The problem is that users want to retrieve on the basis of conceptual content, and individual words provide unreliable evidence about the conceptual topic or meaning of a document. There are usually many ways to express a given concept, so the literal terms in a user's query may not match those of a relevant document. In addition, most words have multiple meanings, so terms in a user's query will literally match terms in documents that are not of interest to the user.

The proposed approach tries to overcome the deficiencies of term-matching retrieval by treating the unreliability of observed term-document association data as a statistical problem. We assume there is some underlying latent semantic structure in the data that is partially obscured by the randomness of word choice with respect to retrieval. We use statistical techniques to estimate this latent structure, and get rid of the obscuring "noise." A description of terms and documents based on the latent semantic structure is used for indexing and retrieval.[1]

The particular "latent semantic indexing" (LSI) analysis that we have tried uses singular-value decomposition. We take a large matrix of term-document association data and construct a "semantic" space wherein terms and documents that are closely associated are placed near one another. Singular-value decomposition allows the arrangement of the space to reflect the major associative patterns in the data, and ignore the smaller, less important influences. As a result, terms that did not actually appear in a document may still end up close to the document, if that is consistent with the major patterns of association in the data. Position in the space then serves as the new kind of semantic indexing. Retrieval proceeds by using the terms in a query to identify a point in the space, and documents in its neighborhood are returned to the user.

Deficiencies of Current Automatic Indexing and Retrieval Methods

A fundamental deficiency of current information retrieval methods is that the words searchers use often are not the same as those by which the information they seek has been indexed. There are actually two sides to the issue; we will call them broadly *synonymy* and *polysemy*. We use *synonymy* in a very general sense to describe the fact that

*To whom all correspondence should be addressed.

Received August 26, 1987; revised April 4, 1988; accepted April 5, 1988.

[1] By "semantic structure" we mean here only the correlation structure in the way in which individual words appear in documents; "semantic" implies only the fact that terms in a document may be taken as referents to the document itself or to its topic.

there are many ways to refer to the same object. Users in different contexts, or with different needs, knowledge, or linguistic habits will describe the same information using different terms. Indeed, we have found that the degree of variability in descriptive term usage is much greater than is commonly suspected. For example, two people choose the same main key word for a single well-known object less than 20% of the time (Furnas, Landauer, Gomez, & Dumais, 1987). Comparably poor agreement has been reported in studies of interindexer consistency (Tarr & Borko, 1974) and in the generation of search terms by either expert intermediaries (Fidel, 1985) or less experienced searchers (Liley, 1954; Bates, 1986). The prevalence of synonyms tends to decrease the "recall" performance of retrieval systems. By *polysemy* we refer to the general fact that most words have more than one distinct meaning (homography). In different contexts or when used by different people the same term (e.g., "chip") takes on varying referential significance. Thus the use of a term in a search query does not necessarily mean that a document containing or labeled by the same term is of interest. Polysemy is one factor underlying poor "precision."

The failure of current automatic indexing to overcome these problems can be largely traced to three factors. The first factor is that the way index terms are identified is incomplete. The terms used to describe or index a document typically contain only a fraction of the terms that users as a group will try to look it up under. This is partly because the documents themselves do not contain all the terms users will apply, and sometimes because term selection procedures intentionally omit many of the terms in a document.

Attempts to deal with the synonymy problem have relied on intellectual or automatic term expansion, or the construction of a thesaurus. These are presumably advantageous for conscientious and knowledgeable searchers who can use such tools to suggest additional search terms. The drawback for fully automatic methods is that some added terms may have different meaning from that intended (the polysemy effect) leading to rapid degradation of precision (Sparck Jones, 1972).

It is worth noting in passing that experiments with small interactive data bases have shown monotonic improvements in recall rate without overall loss of precision as more indexing terms, either taken from the documents or from large samples of actual users' words are added (Gomez, Lochbaum, & Landauer, in press; Furnas, 1985). Whether this "unlimited aliasing" method, which we have described elsewhere, will be effective in very large data bases remains to be determined. Not only is there a potential issue of ambiguity and lack of precision, but the problem of identifying index terms that are not in the text of documents grows cumbersome. This was one of the motives for the approach to be described here.

The second factor is the lack of an adequate automatic method for dealing with polysemy. One common approach is the use of controlled vocabularies and human intermediaries to act as translators. Not only is this solution extremely expensive, but it is not necessarily effective. Another approach is to allow Boolean intersection or coordination with other terms to disambiguate meaning. Success is severely hampered by users' inability to think of appropriate limiting terms if they do exist, and by the fact that such terms may not occur in the documents or may not have been included in the indexing.

The third factor is somewhat more technical, having to do with the way in which current automatic indexing and retrieval systems actually work. In such systems each word type is treated as independent of any other (see, for example, van Rijsbergen (1977)). Thus matching (or not) both of two terms that almost always occur together is counted as heavily as matching two that are rarely found in the same document. Thus the scoring of success, in either straight Boolean or coordination level searches, fails to take redundancy into account, and as a result may distort results to an unknown degree. This problem exacerbates a user's difficulty in using compound-term queries effectively to expand or limit a search.

Rationale of the Latent Semantic Indexing (LSI) Method

Illustration of Retrieval Problems

We illustrate some of the problems with term-based information retrieval systems by means of a fictional matrix of terms by documents (Table 1). Below the table we give a fictional query that might have been passed against this database. An "R" in the column labeled REL (relevant) indicates that the user would have judged the document relevant to the query (here documents 1 and 3 are relevant). Terms occurring in both the query and a document (*computer* and *information*) are indicated by an asterisk in the appropriate cell; an "M" in the MATCH column indicates that the document matches the query and would have been returned to the user. Documents 1 and 2 illustrate common classes of problems with which the proposed method

TABLE 1. Sample term by document matrix.*

	Access	Document	Retrieval	Information	Theory	Database	Indexing	Computer	REL	MATCH
Doc 1	x	x	x			x	x		R	
Doc 2				x*	x			x*		M
Doc 3			x	x*				x*	R	M

*Query: "IDF in *computer-based information look-up*" deals. Document 1 is a relevant document, which, however, contains none of the words in the query. It would, therefore, not be returned by a straightforward term overlap retrieval scheme. Document 2 is a nonrelevant document which does contain terms in the query, and therefore would be returned, despite the fact that the query context makes it clear enough to a human observer that a different sense of at least one of the words is intended. Note that in this example none of the meaning conditioning terms in the query is found in the index. Thus intersecting them with the query terms would not have been a plausible strategy for omitting document 2.

Start by considering the synonymy problem. One way of looking at the problem is that document 1 should have contained the term "look-up" from the user's perspective, or conversely that the query should have contained the term "access" or "retrieval" from the system's. To flesh out the analogy, we might consider any document (or title or abstract) to consist of a small selection from the complete discourse that might have been written on its topic. Thus the text from which we extract index terms is a fallible observation from which to infer what terms actually apply to its topic. The same can be said about the query: it is only one sample description of the intended documents, and in principle could have contained many different terms from the ones it does.

Our job then, in building a retrieval system, is to find some way to predict what terms "really" are implied by a query or apply to a document (i.e., the "latent semantics") on the basis of the fallible sample actually found there. If there were no correlation between the occurrence of one term and another, then there would be no way for us to use the data in a term by document matrix to estimate the "true" association of terms and documents where data are in error. On the other hand, if there is a great deal of structure, i.e., the occurrence of some patterns of words gives us a strong clue as to the likely occurrence of others, then data from one part (or all) of the table can be used to correct other portions. For example suppose that in our total collection the words "access" and "retrieval" each occurred in 100 documents, and that 95 of these documents containing "access" also contained "retrieval." We might reasonably guess that the absence of "retrieval" from a document containing "access" might be erroneous, and consequently wish to retrieve the document in response to a query containing only "retrieval." The kind of structure on which such inferences can be based is not limited to simple pairwise correlation.

In document 2 we would like our analysis to tell us that the term "information" is in fact something of an imposter. Given the other terms in the query and in that document we would predict no occurrence of a term with the meaning here intended for "information," i.e., knowledge desired by a searcher. A correlational structure analysis may allow us to down-weight polysemous terms by taking advantage of such observations.

Our overall research program has been to find effective models for overcoming these problems. We would like a representation in which a set of terms, which by itself is incomplete and unreliable evidence of the relevance of a given document, is replaced by some other set of entities which are more reliable indicants. We take advantage of implicit higher-order (or latent) structure in the association of terms and documents to reveal such relationships.

The Choice of Method for Uncovering Latent Semantic Structure

The goal is to find and fit a useful model of the relationships between terms and documents. We want to use the matrix of observed occurrences of terms applied to documents to estimate parameters of that underlying model. With the resulting model we can then estimate what the observed occurrences really should have been. In this way, for example, we might predict that a given term should be associated with a document, even though, because of variability in word use, no such association was observed.

The first question is what sort of model to choose. A notion of semantic similarity, between documents and between terms, seems central to modeling the patterns of term usage across documents. This led us to restrict consideration to proximity models, i.e., models that try to put similar items near each other in some space or structure. Such models include: hierarchical, partition and overlapping clusterings; ultrametric and additive trees; and factor-analytic and multidimensional distance models (see Carroll & Arabie, 1980 for a survey).

Aiding information retrieval by discovering latent proximity structure has at least two lines of precedence in the literature. Hierarchical classification analyses are frequently used for term and document clustering (Sparck Jones, 1971; Salton, 1968; Jardin & van Rijsbergen, 1971). Latent class analysis (Baker, 1962) and factor analysis (Atherton & Borko, 1965; Borko & Bernick, 1963; Ossorio, 1966) have also been explored before for automatic document indexing and retrieval.

In document clustering, for example, a notion of distance is defined such that two documents are considered close to the extent that they contain the same terms. The matrix of document-to-document distances is then subjected to a clustering analysis to find a hierarchical classification for the documents. Retrieval is based on exploring neighborhoods of this structure. Similar efforts have analyzed word usage in a corpus and built clusters of related terms, in effect making a statistically-based thesaurus. We believe an important weakness of the clustering approach is that hierarchies are far too limited to capture the rich semantics of most document sets. Hierarchical clusterings permit no cross classifications, for example, and in general have very few free parameters (essentially only *n* parameters for *n* objects). Empirically, clustering improves the computational efficiency of search; whether or not it improves retrieval success is unclear (Jardin & van Rijsbergen, 1971; Salton & McGill, 1983; Voorhees, 1985).

Previously tried factor analytic approaches have taken a square symmetric matrix of similarities between pairs of documents (based on statistical term overlap or human judgments), and used linear algebra to construct a low dimensional spatial model wherein similar documents are placed near one another. The factor analytic model has the potential of much greater richness than the clustering model (a *k* dimensional model for *n* points has *nk* parameters). However previous attempts along these lines, too, had shortcomings. First, factor analysis is computationally expensive, and since most previous attempts were made 15–20 years ago, they were limited by processing constraints (Borko & Bernick, 1963). Second, most past attempts considered restricted versions of the factor analytic model, either by using very low dimensionality, or by converting the factor analysis results to a simple binary clustering (Borko & Bernick, 1963). Third, some attempts have relied on excessively tedious data gathering techniques, requiring the collection of thousands of similarity judgments from humans (Ossorio, 1966).

Previously reported clustering and factor analytic approaches have also struggled with a certain representational awkwardness. Typically the original data explicitly relate two types of entities, terms and documents, and most conceptions of the retrieval problem mention both types (e.g., given terms that describe a searchers' interests, relevant documents are returned). However, representations chosen so far handle only one at a time (e.g., either term clustering or document clustering). Any attempts to put the ignored entity back in the representation have been arbitrary and after the fact. An exception to this is a proposal by Koll (1979) in which both terms and documents are represented in the same space of concepts (see also Raghavan & Wong (1986)). While Koll's approach is quite close in spirit to the one we propose, his concept space was of very low dimensionality (only seven underlying dimensions), and the dimensions were hand-chosen and not truly orthogonal as are the underlying axes in factor analytic approaches.[2]

Our approach differs from previous attempts in a number of ways that will become clearer as the model is described in more detail. To foreshadow some of these differences, we: (1) examine problems of reasonable size (1000–2000 document abstracts; and 5000–7000 index terms); (2) use a rich, high-dimensional representation (about 100 dimensions) to capture term-document relations (and this appears necessary for success); (3) use a mathematical technique which explicitly represents both terms and documents in the same space; and (4) retrieve documents from query terms directly, without rotation or interpretation of the underlying axes and without using intermediate document clusters.

We considered alternative models using the following three criteria:

(1) *Adjustable representational richness.* To represent the underlying semantic structure, we need a model with sufficient power. We believe hierarchical clusterings to be too restrictive, since they allow no multiple or crossed classifications and have essentially only as many parameters as objects. Since the right kind of alternative is unknown, we looked for models whose power could be varied, as some compensation for choosing a perhaps inappropriate structure. The most obvious class is dimensional models, like multidimensional scaling and factor analysis, where representational power can be controlled by choosing the number, *k*, of dimensions (i.e., *k* parameters per object).

(2) *Explicit representation of both terms and documents.* The desire to represent both terms and documents simultaneously is more than esthetic. In our proximity-based latent structure paradigm, retrieval proceeds by appropriately placing a new object corresponding to the query in the semantic structure and finding those documents that are close by. One simple way to achieve appropriate placement is if terms, as well as documents, have positions in the structure. Then a query can be placed at the centroid of its term points. Thus for both elegance and retrieval mechanisms, we needed what are called two-mode proximity methods (Carroll and Arabie, 1980), that start with a rectangular matrix and construct explicit representations of both row and column objects. One such method is multidimensional unfolding (Coombs, 1964; Heiser, 1981; Desarbo & Carroll, 1985), in which both terms and documents would appear as points in a single space with similarity related monotonically to Euclidean distance. Another is two-mode factor analysis (Harshman, 1970; Harshman & Lundy, 1984a; Carroll & Chang, 1970; Kruskal, 1978), in which terms and documents would again be represented as points in a space, but similarity is given by the inner product between points. A final candidate is unfolding in trees (Furnas, 1980), in which both terms and documents would appear as leaves on a tree, and path length distance through the tree would give the similarity. (One version of this is equivalent to simultaneous hierarchical clustering of both terms and objects.) The explicit representation of both terms and documents also leads to a straightforward way in which to add or "fold-in" new terms or documents that were not in the original matrix. New terms can be placed at the centroid of the documents in which they appear; similarly, new documents can be placed at the centroid of their constituent terms.[3]

[2] Koll begins with a set of seven nonoverlapping but almost spanning documents which form the axes of the space. Terms are located on the axis of the document in which they occur; the remainder of the documents are processed sequentially and placed at the average of their terms. This approach has been evaluated on only a small dataset where it was moderately successful.

[3] There are several important and interesting issues raised by considering the addition of new terms and documents into the space. First, the addition of new objects introduces some temporal dependencies in the representation. That is, where a new term or document gets placed depends on what other terms and documents are already in the space. Second, in general, simply folding-in new terms or documents will result in a somewhat different space than would have been obtained had these objects been included in the original analysis. Since the initial analysis is time consuming, it is clearly advantageous to be able to add new objects by folding-in. How much of this can be done without rescaling is an open research issue, and is likely to depend on the variability of the database over time, the representativeness of the original of documents and terms, etc.

(3) *Computational tractability for large datasets.* Many of the existing models require computation that goes up with N^4 or N^5 (where N is the number of terms plus documents). Since we hoped to work with documents sets that were at least in the thousands, models with efficient fitting techniques were needed.

The only model which satisfied all three criteria was two-mode factor analysis. The tree unfolding model was considered too representationally restrictive, and along with nonmetric multidimensional unfolding, too computationally expensive. Two-mode factor analysis is a generalization of the familiar factor analytic model based on singular value decomposition (SVD). (See Forsythe, Malcolm, & Moler (1977), Chapter 9, for an introduction to SVD and its applications.) SVD represents both terms and documents as vectors in a space of choosable dimensionality, and the dot product or cosine between points in the space gives their similarity. In addition, a program was available (Harshman & Lundy, 1984b) that fit the model in time of order $N^2 \times k^3$.

SVD or Two-Mode Factor Analysis

Overview

The latent semantic structure analysis starts with a matrix of terms by documents. This matrix is then analyzed by singular value decomposition (SVD) to derive our particular latent semantic structure model. Singular value decomposition is closely related to a number of mathematical and statistical techniques in a wide variety of other fields, including eigenvector decomposition, spectral analysis, and factor analysis. We will use the terminology of factor analysis, since that approach has some precedence in the information retrieval literature.

The traditional, one-mode factor analysis begins with a matrix of associations between all pairs of one type of object, e.g., documents (Borko & Bernick, 1963). This might be a matrix of human judgments of document to document similarity, or a measure of term overlap computed for each pair of documents from an original term by document matrix. This square symmetric matrix is decomposed by a process called "eigen-analysis," into the product of two matrices of a very special form (containing "eigenvectors" and "eigenvalues"). These special matrices show a breakdown of the original data into linearly independent components or "factors." In general many of these components are very small, and may be ignored, leading to an approximate model that contains many fewer factors. Each of the original documents' similarity behavior is now approximated by its values on this smaller number of factors. The result can be represented geometrically by a spatial configuration in which the dot product or cosine between vectors representing two documents corresponds to their estimated similarity.

In two-mode factor analysis one begins not with a square symmetric matrix relating pairs of only one type of entity, but with an arbitrary rectangular matrix with different entities on the rows and columns, e.g., a matrix of terms and documents. This rectangular matrix is again decomposed into three other matrices of a very special form, this time by a process called "singular-value-decomposition" (SVD). (The resulting matrices contain "singular vectors" and "singular values.") As in the one-mode case these special matrices show a breakdown of the original relationships into linearly independent components or factors. Again, many of these components are very small, and may be ignored, leading to an approximate model that contains many fewer dimensions. In this reduced model all the term-term, document-document, and term-document similarities are now approximated by values on this smaller number of dimensions. The result can still be represented geometrically by a spatial configuration in which the dot product or cosine between vectors representing two objects corresponds to their estimated similarity.

Thus, for information retrieval purposes, SVD can be viewed as a technique for deriving a set of uncorrelated indexing variables or factors; each term and document is represented by its vector of factor values. Note that by virtue of the dimension reduction, it is possible for documents with somewhat different profiles of term usage to be mapped into the same vector of factor values. This is just the property we need to accomplish the improvement of unreliable data proposed earlier. Indeed, the SVD representation, by replacing individual terms with derived orthogonal factor values, can help to solve all three of the fundamental problems we have described.

In various problems, we have approximated the original term-document matrix using 50–100 orthogonal factors or derived dimensions. Roughly speaking, these factors may be thought of as artificial concepts; they represent extracted common meaning components of many different words and documents. Each term or document is then characterized by a vector of weights indicating its strength of association with each of these underlying concepts. That is, the "meaning" of a particular term, query, or document can be expressed by k factor values, or equivalently, by the location of a vector in the k-space defined by the factors. The meaning representation is economical, in the sense that N original index terms have been replaced by the $k < N$ best surrogates by which they can be approximated. We make no attempt to interpret the underlying factors, nor to "rotate" them to some meaningful orientation. Our aim is not to be able to describe the factors verbally but merely to be able to represent terms, documents and queries in a way that escapes the unreliability, ambiguity and redundancy of individual terms as descriptors.

It is possible to reconstruct the original term by document matrix from its factor weights with reasonable but not perfect accuracy. It is important for the method that the derived k-dimensional factor space *not* reconstruct the original term space perfectly, because we believe the original term space to be unreliable. Rather we want a derived structure that expresses what is reliable and important in the underlying use of terms as document referents.

Unlike many typical uses of factor analysis, we are not necessarily interested in reducing the representation to a very low dimensionality, say two or three factors, because we are not interested in being able to visualize the space or understand it. But we do wish both to achieve sufficient power and to minimize the degree to which the space is distorted. We believe that the representation of a conceptual space for any large document collection will require more than a handful of underlying independent "concepts," and thus that the number of orthogonal factors that will be needed is likely to be fairly large. Moreover, we believe that the model of a Euclidean space is at best a useful approximation. In reality, conceptual relations among terms and documents certainly involve more complex structures, including, for example, local hierarchies, and nonlinear interactions between meanings. More complex relations can often be made to approximately fit a dimensional representation by increasing the number of dimensions. In effect, different parts of the space will be used for different parts of the language or object domain. Thus we have reason to avoid both very low and extremely high numbers of dimensions. In between we are guided only by what appears to work best. What we mean by "works best" is not (as is customary in some other fields) what reproduces the greatest amount of variance in the original matrix, but what will give the best retrieval effectiveness.

How do we process a query in this representation? Recall that each term and document is represented as a vector in k-dimensional factor space. A query, just as a document, initially appears as a set of words. We can represent a query (or "pseudo-document") as the weighted sum of its component term vectors. (Note that the location of each document can be similarly described; it is a weighted sum of its constituent term vectors.) To return a set of potential candidate documents, the pseudo-document formed from a query is compared against all documents, and those with the highest cosines, that is the nearest vectors, are returned. Generally, either a threshold is set for closeness of documents and all those above it returned, or the n closest are returned. (We are concerned with the issue of whether the cosine measure is the best indication of similarity to predict human relevance judgments, but we have not yet systematically explored any alternatives, cf. Jones and Furnas, 1987.)

A concrete example may make the procedure and its putative advantages clearer. Table 2 gives a sample dataset. In this case, the document set consisted of the titles of nine Bellcore technical memoranda. Words occurring in more than one title were selected for indexing; they are italicized. Note that there are two classes of titles: five about human-computer interaction (labeled c1–c5) and four about graph theory (labeled m1–m4). The entries in the term by document matrix are simply the frequencies with which each term actually occurred in each document. Such a matrix could be used directly for keyword-based retrievals or, as here, for the initial input of the SVD analysis. For this example we carefully chose documents and terms so that SVD would produce a satisfactory solution using just two dimensions. Figure 1 shows the two-dimensional geometric representation for terms and documents that resulted from the SVD analysis. Details of the mathematics underlying the analysis will be presented in the next section. The numerical results of the SVD for this example are shown in the appendix and can be used to verify the placement of terms and documents in Figure 1. Terms are shown as filled circles and labeled accordingly; document titles are represented by open squares, with the numbers of the terms contained in them indicated parenthetically. Thus, each term and document can be described by its position in this two-dimensional factor space.

TABLE 2. A sample dataset consisting of the titles of nine technical memoranda. Terms occurring in more than one title are italicized. There are two classes of documents — five about human-computer interaction (c1-c5) and four about graphs (m1-m4). This dataset can be described by means of a term by document matrix where each cell entry indicates the frequency with which a term occurs in a document.

Technical Memo Example

Titles	
c1	*Human* machine *interface* for Lab ABC *computer* applications
c2	A *survey* of *user* opinion of *computer system response time*
c3	The *EPS user interface* management *system*
c4	*System* and *human system* engineering testing of *EPS*
c5	Relation of *user*-perceived *response time* to error measurement
m1	The generation of random, binary, unordered *trees*
m2	The intersection *graph* of paths in *trees*
m3	*Graph minors* IV: Widths of *trees* and well-quasi-ordering
m4	*Graph minors*: A *survey*

Terms	Documents								
	c1	c2	c3	c4	c5	m1	m2	m3	m4
human	1	0	0	1	0	0	0	0	0
interface	1	0	1	0	0	0	0	0	0
computer	1	1	0	0	0	0	0	0	0
user	0	1	1	0	1	0	0	0	0
system	0	1	1	2	0	0	0	0	0
response	0	1	0	0	1	0	0	0	0
time	0	1	0	0	1	0	0	0	0
EPS	0	0	1	[illegible]	0	0	0	0	0
survey	0	1	0	0	0	0	0	0	1
trees	0	0	0	0	0	1	1	1	0
graph	0	0	0	0	0	0	1	1	1
minors	0	0	0	0	0	0	0	1	1

One test we set ourselves is to find documents relevant to the query: "human computer interaction." Simple term matching techniques would return documents c1, c2, and c4 since they share one or more terms with the query. However, two other documents, which are also relevant (c3 and c5), are missed by this method since they have no terms in common with the query. The latent semantic structure method uses the derived factor representation to process the query; the first two-dimensions are shown in Figure 1. First, the query is represented as a "pseudo-document" in the factor space. Two of the query terms, "human" and "computer," are in the factor space, so the query is placed at their centroid and scaled for comparison to documents (the point labeled q in Figure 1 represents the query).

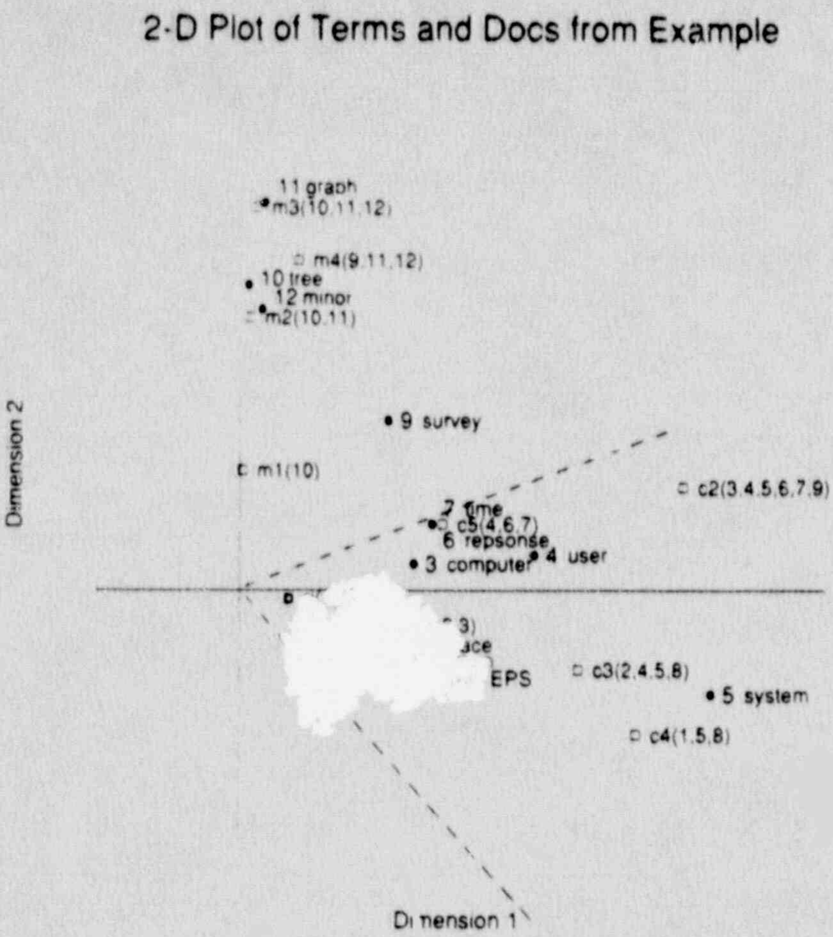


FIG 1 A two-dimensional plot of 12 Terms and 9 Documents from the sai ıpe TM set. Terms ar. represented by filled circles. Documents are shown as open squares, and component terms are indicated parenthetically. The que ry ("human computer interaction") is represented as a pseudo-document at point q. Axes are scaled for Document-Document or Term-Term comparisor s. The dotted cone represents the region whose points are within a cosine of .9 from the query q. All documents about human-computer (c1–c5) are "n ar" the query (i.e., within this cone), but none of the graph theory documents (m1–m4) are nearby. In this reduced space, even documents c3 and c5 which share no terms with the query are near it.

Then, we simply look for documents which are n‹ar the query, q. In this case, documents c1–c5 (but not m1–m4) are "nearby" (within a cosine of .9, as indicated by the dashed lines). Notice that even documents c3 and c5 which share no index terms at all with the query are near it in this representation. The relations among the documents expressed in the factor space depend on complex and indirect associations between terms and documents, ones that come from an analysis of the structure of the whole set of relations in the term by document matrix. This is the strength of using higher order structure in the term by document matrix to represent the underlying meaning of a single term, document, or query. It yields a more robust and economical representation than do term overlap or surface-level clustering methods.

Technical Details

The Singular Value Decomposition (SVD) Model. This section details the mathematics underlying the particular model of latent structure, singular value decomposition, that we currently use. The casual reader may wish to skip this section and proceed to the next section.

Any rectangular matrix, for example a $t \times d$ matrix of terms and documents, X, can be decomposed into the product of three other matrices:

$$X = T_0 S_0 D_0' ,$$

such that T_0 and D_0 have orthonormal columns and S_0 is diagonal. This is called the *singular value decomposition of* X. T_0 and D_0 are the matrices of *left* and *right singular vectors* and S_0 is the diagonal matrix of *singular values*.[4] Singular value decomposition (SVD) is unique up to certain row, column and sign permutations[5] and by convention the diagonal elements of S_0 are constructed to be all positive and ordered in decreasing magnitude.

[4] SVD is closely related to the standard eigenvalue-eigenvector or spectral decomposition of a square symmetric matrix, Y, into VLV', where V is orthonormal and L is diagonal. The relation between SVD and eigen analysis is more than one of analogy. In fact, T_0 is the matrix of eigenvectors of the square symmetric matrix $Y = XX'$, D_0 is th• matrix of eigenvectors of $Y = X'X$, and in both cases, S_0^2 would be the n atrix, L, of eigenvalues.

[5] Allowable permutations are those that leave S_0 diagonal and maintain the correspondences with T_0 and D_0. That is, column i and j of S_0 may be interchanged iff row i and j of S_0 are interchanged, and columns i and j of T_0 and D_0 are interchanged.

Figure 2 presents a schematic of the singular value decomposition for a $t \times d$ matrix of terms by documents.

In general, for $X = T_0 S_0 D_0'$ the matrices T_0, D_0, and S_0 must all be of full rank. The beauty of an SVD, however, is that it allows a simple strategy for optimal approximate fit using smaller matrices. If the singular values in S_0 are ordered by size, the first k largest may be kept and the remaining smaller ones set to zero. The product of the resulting matrices is a matrix $\hat{X}$ which is only approximately equal to X, and is of rank k. It can be shown that the new matrix $\hat{X}$ is the matrix of rank k, which is closest in the least squares sense to X. Since zeros were introduced into S_0, the representation can be simplified by deleting the zero rows and columns of S_0 to obtain a new diagonal matrix S, and then deleting the corresponding columns of T_0 and D_0 to obtain T and D respectively. The result is a reduced model:

$$X \approx \hat{X} = TSD'$$

which is the rank-k model with the best possible least-squares-fit to X. It is this reduced model, presented in Figure 3, that we use to approximate our data.

The amount of dimension reduction, i.e., the choice of k, is critical to our work. Ideally, we want a value of k that is large enough to fit all the real structure in the data, but small enough so that we do not also fit the sampling error or unimportant details. The proper way to make such choices is an open issue in the factor analytic literature. In practice, we currently use an operational criterion—a value of k which yields good retrieval performance.

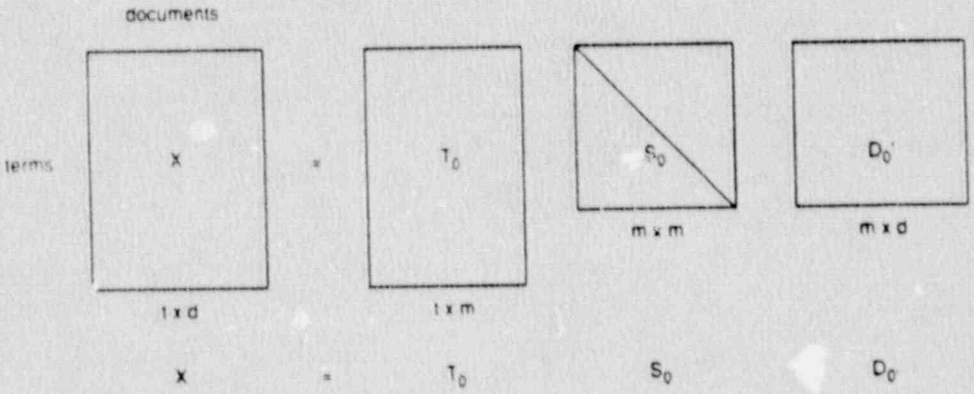
Geometric Interpretation of the SVD Model. For purposes of intuition and discussion it is useful to interpret the SVD geometrically. The rows of the reduced matrices of singular vectors are taken as coordinates of points representing the documents and terms in a k dimensional space. With appropriate rescaling of the axes, by quantities related to the associated diagonal values of S, dot products between points in the space can be used to compare the corresponding objects. The next section details these comparisons.

Computing Fundamental Comparison Quantities from the SVD Model. There are basically three sorts of comparisons of interest: those comparing two terms ("How similar are terms i and j?"), those comparing two documents ("How similar are documents i and j?"), and those comparing a term and a document ("How associated are term i and document j?"). In standard information retrieval approaches, these amount respectively, to comparing two rows, comparing two columns, or examining individual cells of the original matrix of term by document data, X. Here we make similar comparisons, but use the matrix $\hat{X}$, since it is presumed to represent the important and reliable patterns underlying the data in X. Since $\hat{X} = TSD'$, the relevant quantities can be computed just using the smaller matrices, T, D, and S.

Comparing Two Terms. The dot product between two row vectors of $\hat{X}$ reflects the extent to which two terms have a similar pattern of occurrence across the set of documents. The matrix $\hat{X}\hat{X}'$ is the square symmetric matrix containing all these term-to-term dot products. Since S is diagonal and D is orthonormal, it is easy to verify that:

$$\hat{X}\hat{X}' = TS^2T'$$

Note that this means that the i,j cell of $\hat{X}\hat{X}'$ can be obtained by taking the dot product between the i and j rows



Singular value decomposition of the term x document matrix, X. Where:

T_0 has orthogonal, unit-length columns ($T_0' T_0 = I$)
D_0 has orthogonal, unit-length columns ($D_0' D_0 = I$)
S_0 is the diagonal matrix of singular values
t is the number of rows of X
d is the number of columns of X
m is the rank of X ($\leq$ min(t,d))

FIG. 2. Schematic of the Singular Value Decomposition (SVD) of a rectangular term by document matrix. The original term by document matrix is decomposed into three matrices each with linearly independent components.

of the matrix TS. That is, if one considers the rows of TS as coordinates for terms, dot products between these points give the comparison between terms. Note that the relation between taking T as coordinates and taking TS as coordinates is simple since S is diagonal; the positions of the points are the same except that each of the axes has been stretched or shrunk in proportion to the corresponding diagonal element of S.

Comparing Two Documents. The analysis for comparing two documents is similar, except that in this case it is the dot product between two *column* vectors of the matrix $\hat{X}$ which tells the extent to which two documents have a similar profile of terms. Thus the matrix $\hat{X}'\hat{X}$ contains the document-to-document dot products. The definitions of the matrices T, S, and D again guarantee:

$$\hat{X}'\hat{X} = DS^2D'$$

Here the i,j cell of $\hat{X}'\hat{X}$ is obtained by taking the dot product between the i and j rows of the matrix DS. So one can consider rows of a DS matrix as coordinates for documents, and take dot products in this space. (Again note that the DS space is just a stretched version of the D space.)

Comparing a Term and a Document. This comparison is different. Instead of trying to estimate the dot product between rows or between columns of $\hat{X}$, the fundamental comparison between a term and a document is the value of an individual cell of $\hat{X}$. $\hat{X}$ is defined in terms of matrices T, S, and D. Repeating it here:

$$\hat{X} = TSD'$$

The i,j cell of $\hat{X}$ is therefore obtained by taking the dot product between the ith row of the matrix $TS^{1/2}$ and the jth row of the matrix $DS^{1/2}$. Note that while the within comparisons (i.e., term-term or document-document) involve using rows of TS and DS for coordinates, the between comparision requires $TS^{1/2}$ and $DS^{1/2}$ for coordinates. That is, it is not possible to make a single configuration of points in a space that will allow both between and within comparisions. They will be similar however, differing only by a stretching or shrinking of the axes by a factor of $S^{1/2}$.

Finding Representations for Pseudo-Documents. The previous results show how it is possible to compute comparisions between the various objects associated with the rows or columns of $\hat{X}$. It is very important in information retrieval applications to compute appropriate comparison quantities for objects that did not appear in the original analysis. For example, we want to be able to take a completely novel query, find some point for it in the space, and then look at its cosine with respect to terms or documents in the space. Another example would be trying, after-the-fact, to find representations for documents that did not appear in the original analysis. The new objects in both these examples are very much like the documents of the matrices, X and $\hat{X}$, in that they present themselves as vectors of terms. It is for this reason that we call them *pseudo-documents*. In order to compare a query or pseudo-document, q, to other documents, we need to be able to start with its term vector X_q and derive a representation D_q that we can use just like a row of D in the comparison formulas of the preceding section. One criterion for such a derivation is that putting in a real document X_i should give D_i (at least when the model is perfect, i.e., $X = \hat{X}$). With this constraint, a little algebra shows that:

$$D_q = X_q'TS^{-1}$$

Note that with appropriate rescaling of the axes, this amounts to placing the pseudo-document at the centroid of its corresponding term points. This D_q then is just like a row of D and, appropriately scaled by $S^{1/2}$ or S, can be used like a usual document's factor vector for making between or within comparisons, respectively.

Preprocessing and Normalization. The equations given here do not take into account any preprocessing or reweighting of the rows or columns of X. Such preprocessing might be used to prevent documents of different overall length from having differential effect on the model, or be used to impose certain preconceptions of which terms are more important. The effects of certain of these transformations can be taken into account in a straightforward way, but we will not go into the algebra here.

Tests of the SVD Latent Semantic Indexing (LSI) Method

We have so far tried the LSI method on two standard document collections where queries and relevance judgments were available (MED and CISI). PARAFAC (Harshman & Lundy, 1984b), a program for the iterative numerical solution of multi-mode factor-analysis problems, was used for the studies reported below. (Other programs for more standard SVD are also available—e.g., Golub, Luk, and Overton, 1981; Cullum, Willoughby, and Lake, 1983.)

"Documents" consist of the full text of the title and abstract. Each document is indexed automatically; all terms occurring in more than one document and not on a stop list of 439 common words used by SMART are included in the analyses.[6] We did *not* stem words or map variants of words to the same root form. The analysis begins with a term by document matrix in which each cell indicates the frequency with which each term occurs in each document. This matrix was analyzed by singular value decomposition to derive our latent structure model which was then used for indexing and retrieval. Queries were placed in the resulting space at the centroid of their constituent terms (again, all terms not on the stop list and occurring in more

[6] We have argued above that the more terms the better, but so far, computational constraints have limited us to around 7000 terms. Terms that occur in only one document, or equally frequently in all documents have little or no influence on the SVD solution. Rejecting such terms has usually been sufficient to satisfy our computational constraints. (In addition, we wanted to be as consistent with SMART as possible in indexing, thus the omission of SMART's common words.) Given greater resources, we see no reason to omit any terms from the latent structure analysis. Even given current limited computational resources, the terms omitted in indexing can be used for *retrieval* purposes by folding them back into the concept space, as we described briefly in the text.

than one document were used). Cosines between the query vector and document vectors are then straightforward to compute (see the "Technical Details" section for details), and documents are ordered by their distance to the query. In many senses, the current LSI method is impoverished and thus provides a conservative test of the utility of latent semantic structure in indexing and information retrieval. We have so far avoided adding refinements such as stemming, phrases, term-weighting, and Boolean combinations (all of which generally result in performance improvements) in order to better evaluate the utility of the basic representation technique.

We compare the results of our latent structure indexing (LSI) method against a straightforward term matching method, a version of SMART, and against data reported by Voorhees (1985) for the same standard datasets. The term overlap comparisons provide a baseline against which to assess the benefits of indexing by means of latent semantic structure rather than raw term matching. For the term matching method, we use the same term-document matrix that was the starting point for the LSI method. A query is represented as a column, and cosines between the query column and each document column are calculated. The SMART and Voorhees systems are more representative of state of the art information retrieval systems, but differences in indexing, term weighting, and query processing preclude precise comparisons of our LSI method and these systems. Nonetheless, such comparisons are of interest. For the SMART evaluations, documents were indexed using a stop list of common words, full stemming, and raw term frequencies as options. Queries were similarly processed and a vector sequential search was used for matching queries and documents. This particular invocation of SMART is the same as our term matching method except for the initial choice of index terms. The Voorhees data were obtained directly from her paper in which she used a vector retrieval system with extended Boolean queries (see Voorhees (1985) for details). Her documents were indexed by removing words on a stop list, mapping word variants into the same term, and weighting terms. Weighted extended Boolean queries were used for retrieval.

Performance is evaluated by measuring precision at several different levels of recall. This is done separately for each available query and then averaged over queries. For the LSI, term matching, and SMART runs, full precision-recall curves can be calculated. For the Voorhees data, only two precision-recall pairs are available; these are the values obtained when 10 or 20 documents were returned—see her Figures 4b and 6b. We present the values from her sequential search (SEQ) condition, since the best performance is generally observed in this condition and not in one of the retrieval conditions using document clusters.

MED

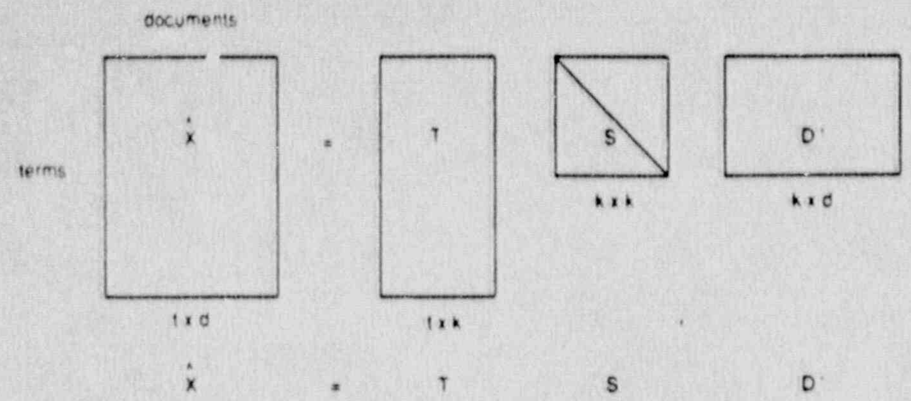
The first standard set we tried, MED, was the commonly studied collection of medical abstracts. It consists of 1033 documents and 30 queries. Our automatic indexing on all terms occurring in more than one document and not on SMART's stop list of common words resulted in 5823 indexing terms. Some additional characteristics of the dataset are given below:

	Term and LSI	Voorhees	SMART
Number of unique terms	5823	6927	6927
Mean number of terms per document	50.1	51.6	51.6
Mean number of terms per query	9.8	39.7[7]	10.1
Mean number of relevant documents per query	23.2	23.2	23.2

The number of unique terms, terms per document, and terms per query vary somewhat because different term-processing algorithms were used in the different systems. A 100-factor SVD of the 5823 term by 1033 document matrix was obtained, and retrieval effectiveness evaluated against the 30 queries available with the dataset. Figure 4 shows precision as a function of recall for a LSI 100-factor solution ("LSI-100"), term matching ("TERM"), SMART ("SMART"), and the Voorhees data ("VO"), all on the same set of documents and queries.

For all but the two lowest levels of recall (.10), precision of the LSI method lies well above that obtained with straightforward term matching, SMART, and the vector method reported by Voorhees. The average difference in precision between the LSI and the term matching method is .06 (.51 vs. .45), which represents a 13% improvement over raw term matching. (The odds against a difference this large or larger by chance is 29 to 1, $t(29) = 2.23$.) Thus, LSI captures some structure in the data which is obscured when raw term overlap is used. The LSI method also compares favorably with SMART ($t(29) = 1.96$; the odds against a difference this large or larger by chance is 16 to 1) and the Voorhees system. It is somewhat surprising that the term matching and SMART methods do not differ for this data set. There are several differences in indexing between LSI and SMART (word stemming is used in SMART but not LSI, and SMART includes word stems occurring in any document whereas LSI, for computational reasons, includes only terms occurring in more than one document) that should lead to better performance for SMART. The difference in performance between LSI and the other methods is especially impressive at higher recall levels where precision is ordinarily quite low, thus representing large proportional improvements. The comparatively poor performance of the LSI method at the lowest levels of recall can be traced to at least two factors. First, precision is quite good in all systems at low recall, leaving little room for improvement. Second, latent semantic indexing is designed primarily to handle synonymy problems (thus improving recall); it is less successful in dealing with polysemy (precision). Synonymy is not much of a problem at low recall since any word matches will retrieve some of

[7] The value 39.7 is reported in Table 1 of the Voorhees article. We suspect this is in error, and that the correct value may be 9.7 which would be in line with the other measures of mean number of terms per query.



Reduced singular value decomposition of the term x document matrix, X. Where:

T has orthogonal, unit-length columns (T' T = I)
D has orthogonal, unit-length columns (D' D = I)
S is the diagonal matrix of singular values

t is the number of rows of X
d is the number of columns of X
m is the rank of X ($\leq$ min(t,d))
k is the chosen number of dimensions in the reduced model ($k \leq m$)

FIG. 3. Schematic of the *reduced* Singular Value Decomposition (SVD) of a term by document matrix. The original term by document matrix is *approximated* using the *k* largest singular values and their corresponding singular vectors.

the relevant documents. Thus the largest benefits of the LSI method should be observed at high recall, and, indeed, this is the case.

Up to this point, we have reported LSI results from a 100-factor representation (i.e., in a 100-dimensional space). This raises the important issue of choosing the dimension-

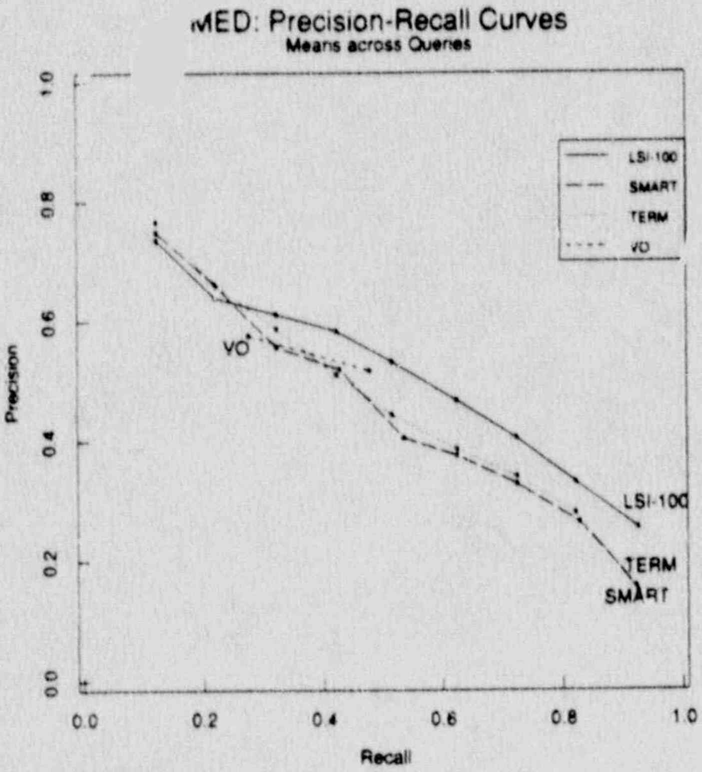


FIG. 4. Precision-recall curves for TERM matching, a 100-factor LSI, SMART, and Voorhees systems on the MED dataset. The data for the TERM matching, LSI, and SMART methods are obtained by measuring precision at each of nine levels of recall (approximately .10 increments) for each query separately and then averaging over queries. The two Voorhees data points are taken from Table 4b in her article.

ality. Ideally we want enough dimensions to capture all the real structure in the term-document matrix, but not too many, or we may start modeling noise or irrelevant detail in the data. How to choose the appropriate number of dimensions is an open research issue. In our tests, we have been guided by the operational criterion of "what works best." That is, we examine performance for several different numbers of factors, and select the dimensionality which maximizes retrieval performance.[8] The results for the MED dataset are shown in Figure 5 which presents average precision as a function of number of factors. As can be seen, mean precision more than doubles (from .25 to .52) as the number of factors increases from 10 to 100, with a maximum at 100. We therefore use the 100-factor space for the results we report. In this particular dataset, performance might improve a bit if solutions with more than 100 factors were explored, but, in general, it is not the case that more factors necessarily means better performance. (In other small applications, we have seen much clearer maxima; performance increases up to some point, and then decreases when too many factors are used. One interpretation of this decrease is that the extra parameters are modeling the sampling noise or peculiarities of the sample rather than important underlying relationships in the pattern of term usage over documents.) It is also important to note that previous attempts to use factor analytic techniques for information retrieval have used small numbers of factors (Koll (1979), seven dimensions; and Ossorio (1966), 13 dimensions; Borko & Bernick (1963), 21 dimensions). We show a more than 50% improvement in performance beyond this range, and therefore suspect that some of the limited utility of previous factor analytic approaches may be the result of an impoverished representation.

Unfortunately this MED dataset was specially constructed in a way that may have resulted in unrealistically good results. From what we can determine, the test collection was made up by taking the union of the returns of a set of thorough keyword searches for documents relevant to the 30 queries in the set. It thus may be an unrepresentatively well-segmented collection. The sets of documents for particular queries are probably isolated to an abnormal extent in the multidimensional manifold of concepts. In such a circumstance our method does an excellent job of defining the isolated subdomains and separating them for retrieval. This is probably not the way most natural document collections are structured. It is worth noting, however, that other automatic techniques applied to the same dataset are not able to capitalize as well on this same abnormal structural property. Thus the fact that LSI greatly outperforms the rest is still quite significant. (Note also that the use of keyword searches to define the document test set probably biases results in favor of methods based on surface term matching, such as SMART, since no documents that do not contain any of the keywords are included.)

[8] This is actually quite easy to do since the SVD solutions are nested. To explore performance in a 10-dimensional solution, for example, cosines are calculated using only the first 10 coordinates of the 100-factor solution

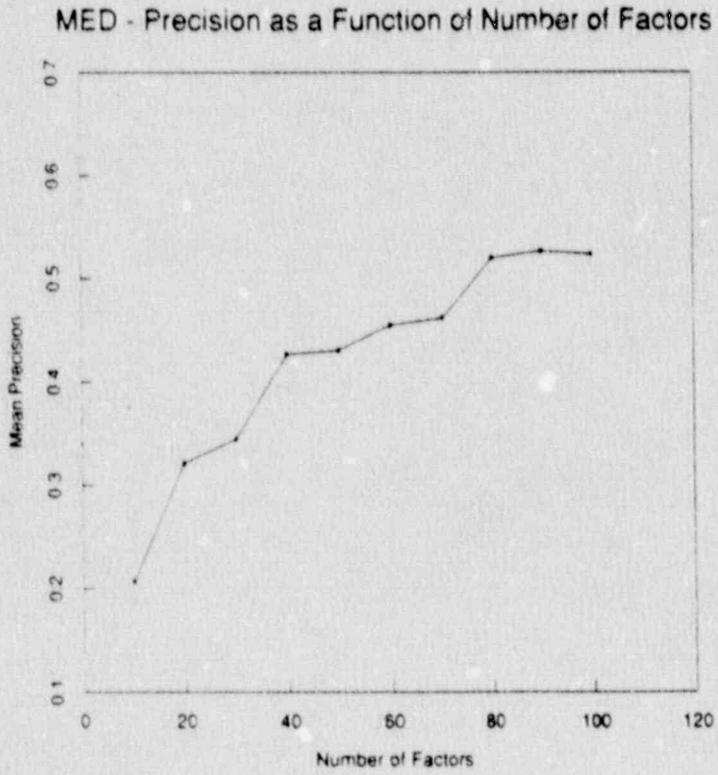


FIG. 5. A plot of average precision (averaged over nine levels of recall) as a function of number of factors for the MED dataset. Precision more than doubles (from about .20 to .50) as the number of factors is increased from 10 to 100.

CISI

Our second test case is the CISI set of 1460 information science abstracts. This set has been consistently difficult for automatic retrieval methods. It consists of 1460 documents and 35 queries. Our automatic indexing, which excluded words on SMART's stop list of common words and words occurring in only one document, resulted in 5135 index terms. Some additional characteristics of the dataset are given below:

	Term and LSI	Voorhees	SMART
Number of unique terms	5135	4941	5019
Mean number of terms per document	45.4	43.9	45.2
Mean number of terms per query	7.7	7.2	7.8
Mean number of relevant documents per query	49.8	49.8	49.8

A 100-factor SVD solution was obtained from the 5135 term by 1460 document matrix, and evaluated using the first 35 queries available with the dataset. LSI results for a 100-factor solution ("LSI-100") along with those for term matching ("TERM"), SMART ("SMART"), and Voorhees ("VO") are shown in Figure 6. All the methods do quite poorly on this dataset, with precision never rising above .30, even for the lowest levels of recall. Average precision is .11 for both LSI and term matching ($t = 1$). For this data set, the latent structure captured by the SVD analysis is no more useful than raw term overlap in capturing the distinctions between relevant and irrelevant documents for the available queries. The Voorhees data cover only a very limited range of low recall levels, but for these values precision is similar to that for LSI and term matching. SMART, on the other hand, results in reliably better performance than LSI, although the absolute levels of precision (.14) is still very low. (The odds against differences this large or larger by chance is over 1000 to 1; $t(34) = 3.66$.) We believe that the superiority of SMART over LSI can be traced to differences in term selection that tend to improve performance. As noted previously, SMART used stemmed words but LSI did not, and SMART included all terms whereas the LSI included only those appearing in more than one document. Since few terms which appear in only one document (and were thus excluded by LSI) are used in the queries, the omission of these words is unlikely to be a major determinant of performance. Thus, stemming appears to be the likely source of performance differences.

We have recently completed a new LSI analysis using SMART's index terms. This enabled us to explore how much of the difference between SMART and the original LSI was due differences in term selection and further to see if additional latent structure could be extracted. For this analysis, we began with a 5019 term (SMART's terms) by 1460 document matrix and obtained a 100-factor SVD solution. The 35 test queries were reevaluated using this new LSI solution (which we refer to as LSI-SMART). The re-

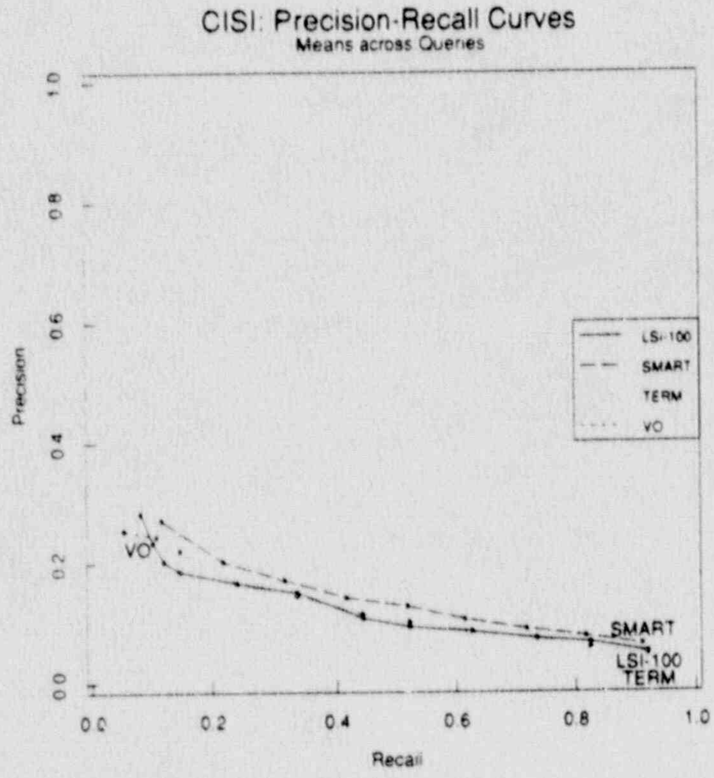


FIG. 6. Precision-recall curves for TERM matching, a 100-factor LSI, SMART, and Voorhees systems on the CISI dataset. The data for TERM matching, LSI, and SMART are obtained by measuring precision at each of nine levels of recall (approximately .10 increments) for each query separately and then averaging over queries. The two Voorhees data points are taken from Table 6b in her paper.

sulting performance was indistinguishable from SMART's: average precision for both methods was .14($t < 1$). This suggests that much of the initial difference between LSI and SMART was due to term selection differences. Unfortunately, LSI was unable to improve upon term matching — either in the initial LSI vs. term matching comparison, or in the LSI-SMART vs. SMART comparison. Stemming, however, seems to capture some structure that LSI was unable to capture as evidenced by the superior performance of SMART relative to term matching. In theory, latent semantic analyses can extract at least some of the commonalities is usage of stemmed forms. In practice, we may often have insufficient data to do so.

A problem in evaluating the CISI dataset is the very low level of precision. Our intuition is that this database contains a very homogeneous distribution of documents that is hard to differentiate on the basis of abstracts. Moreover, many of the test queries, which were given in natural language, seem very vague and poorly stated. Thus the relevance judgments may not be sufficiently reliable to allow any retrieval system to perform well, or to provide an adequate comparison between methods. No direct evidence has been reported on the reliability (repeatability) of these relevance judgments, so this is mostly conjecture, although we do find many cases in which the relevance judgments appear to be in obvious error. In addition, it seems to us that poorly stated queries would invite excessive reliance on term overlap in judging relevance, especially if the judges were familiar with term matching as a possible retrieval strategy.

Summary of Results from LSI Analyses

These results are modestly encouraging. They show the latent semantic indexing method to be superior to simple term matching in one standard case and equal in another. Further, for these two databases, performance with LSI is superior to that obtained with the system described by Voorhees; it performed better than SMART in one case and equal in the other (when term selection differences were eliminated). In order to assess the value of the basic representational method, we have so far avoided the addition of refinements that one would consider in a realistic application, such as discriminative term weighting, stemming, phrase finding or a method of handling negation or disjunction in the queries. So far we have tested the method only with queries formulated to be used against other retrieval methods; the method almost certainly could do better with queries in some more appropriate format. We have projects in progress to add standard enhancements and to incorporate them in a fully automatic indexing and retrieval system. In addition, we are working on methods to incorporate the very low frequency, but often highly informative, words that were filtered out in the trial analysis procedures. It seems likely that with such improvements LSI will offer a more effective retrieval method than has previously been available.

Conclusions and Discussion

Although factor analytic approaches have been previously suggested and tried in the literature, they have all had what we believe to be serious shortcomings which the present attempt overcomes. We have examined problems of reasonable size (1000–2000 document abstracts; and 5000–7000 index terms) using a rich, high-dimensional representation, which appears necessary for success. The explicit representation of both terms and documents in the same space makes retrieving documents relevant to user queries a straightforward matter. Previous work by Borko and his colleagues (Atherton & Borko, 1965; Borko & Bernick, 1963) is similar in name to our approach, but used the factor space only for document clustering, not document retrieval, and computational simplifications reduced its representational power. In Borko and Bernick (1963), for example, factor analysis was performed on a term-term correlation matrix (calculated from word usage over 260 abstracts), and 21 orthogonal factors were selected on the basis of their interpretability. Documents were classified into these 21 categories on the basis of normalized factor loadings for each term in the abstract, and performance was comparable to that of another automatic system. It should be noted, however, that the information used for classification is much less than that which is available in the 21-dimensional factor space, since only the factor loading of "significant" terms on each of the factors was used (e.g., one value for 5, 4, and 7 terms defining the three sample factors presented in their Appendix B). In addition, Borko's work addressed the problem of document classification, and not document retrieval. There is, for example, no discussion of how one might use the full factor space (and not just the document clusters derived from it) for document retrieval.

Koll's (1979) work on concept-based information retrieval is very similar in spirit to our latent semantic indexing. Both terms and documents are represented in a single concept space on the basis of statistical term co-occurrences. Beginning with axes defined by a set of seven nonoverlapping (in terms) and almost-spanning documents, terms were placed on the appropriate axis. New documents were placed at the mean of constituent terms, and new terms were placed at the location of the document in which they occurred. The system was evaluated with only a very small database of documents and queries, but under some circumstances performance was comparable to that of SIRE for Boolean and natural language queries. Our experience with the MED dataset suggests that better performance might have been obtained with a higher dimensional representation. In addition, the latent semantic approach is not order-dependent (as is Koll's procedure), and it is a mathematically rigorous way of uncovering truly orthogonal basis axes or factors for indexing.

The representation of documents by LSI is economical; each document and term need be represented only by something on the order of 50 to 150 values. We have not explored the degree of accuracy needed in these numbers, but we guess that a small integer will probably suffice. The storage requirements for a large document collection can be reduced because much of the redundancy in the characterization of documents by terms is removed in the representation. Offsetting the storage advantage is the fact that the only way documents can be retrieved is by an exhaustive comparison of a query vector against all stored document vectors. Since search algorithms in high dimensional space are not very efficient on serial computers, this may detract from the desirability of the method for very large collections. An additional drawback involves updating. The initial SVD analysis is time consuming, so we would like a more efficient method of adding new terms and documents. We suggest that new documents be located at the centroid of their terms (appropriately scaled); and new terms be placed at the centroid of the documents in which they appear (appropriately scaled). How much of this updating can be done without having to perform a new decomposition is unknown.

While the LSI method deals nicely with the synonymy problem, it offers only a partial solution to the polysemy problem. It helps with multiple meanings because the meaning of a word can be conditioned not only by other words in the document but by other appropriate words in the query not used by the author of a particular relevant document. The failure comes in the fact that every term is represented as just one point in the space. That is, a word with more than one entirely different meaning (e.g., "bank"), is represented as a weighted average of the different meanings. If none of the real meanings is like the average meaning, this may create a serious distortion. (In classical term overlap methods, the meaning of the term is the union of all of it's meanings, which probably leads to less outright distortion, but to more imprecision.) What is needed is some way to detect the fact that a particular term has several distinct meanings and to subcategorize it and place it in several points in the space. We have not yet found a satisfactory way to do that (but see Amsler (1984), Choueka and Lusignan (1985); Lesk (1986))

The latent semantic indexing methods that we have discussed, and in particular the singular-value decomposition technique that we have tested, are capable of improving the way in which we deal with the problem of multiple terms referring to the same object. They replace individual terms as the descriptors of documents by independent "artificial concepts" that can be specified by any one of several terms (or documents) or combinations thereof. In this way relevant documents that do not contain the terms of the query, or whose contained terms are qualified by other terms in the query or document but not both, can be properly characterized and identified. The method yields a retrieval scheme in which documents are ordered continuously by similarity to the query, so that a threshold can be set depending on the desires and resources of the user and service.

At this point in its development, the method should be regarded as a potential component of a retrieval system, rather than as a complete retrieval system as such. As a component it would serve much the same function as is served by raw term vector ranking and other comparison methods. It's putative advantages would be the noise reduction, as described above, and data compaction through the elimination of redundancy. In applying the method, some of the same implementation issues will arise as in raw vector methods — in particular questions of term weighting, stemming, phrasal entries, similarity measure, and counterparts for Boolean operators. Unfortunately, the value of such retrieval enhancing procedures will have to be re-evaluated for use with LSI because its representation changes the nature of the problems with which these procedures were intended to deal. For example, stemming is done to capture likely synonyms. Since LSI already deals with this problem to some extent, the additional value of stemming is an open question. Likewise, LSI averages the "meaning" of polysemous words, where raw term matching maintains one-to-many mappings; as a result, phrases, and other disambiguation techniques may be more important.

Appendix. SVD Numerical Example

In the "Technical Details" section, we outlined the details of the Singular Value Decomposition (SVD) Model. This appendix presents a numerical example using the sample term by document matrix described in the "Overview" section and shown in Table 2 and Figure 1.

The example 12-term by nine-document matrix from Table 2 is presented below.

$$X = \begin{matrix}
1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\
1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\
1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 1 & 1 & 0 & 1 & 0 & 0 & 0 & 0 \\
0 & 1 & 1 & 2 & 0 & 0 & 0 & 0 & 0 \\
0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\
0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\
0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\
0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\
0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1
\end{matrix}$$

Recall that any rectangular matrix, for example a $t \times d$ matrix of terms and documents, X, can be decomposed into the product of three other matrices:

$$X = T_0 S_0 D_0',$$

such that T_0 and D_0 have orthonormal columns and S_0 is diagonal. This is called the *singular value decomposition (SVD)* of X.

Computing the SVD of the X matrix presented above results in the following three matrices for T_0, S_0, D_0 (rounded to two decimal places).

T_0 (nine-dimensional left-singular vectors for 12 terms)
S_0 (diagonal matrix of nine singular values)
D_0 (nine-dimensional right-singular vectors for nine documents)

[redacted]

T_0 =

$$\begin{array}{rrrrrrrrr}
0.22 & -0.11 & 0.29 & -0.41 & -0.11 & -0.34 & 0.52 & -0.06 & -0.41 \\
0.20 & -0.07 & 0.14 & \text{[illegible]} & 0.28 & 0.50 & -0.07 & -0.01 & -0.11 \\
0.24 & 0.04 & -0.16 & \text{[illegible]} & -0.11 & -0.25 & -0.30 & 0.06 & 0.49 \\
0.40 & 0.06 & -0.34 & \text{[illegible]} & 0.33 & 0.38 & 0.00 & 0.00 & 0.01 \\
0.64 & -0.17 & 0.36 & \text{[illegible]} & 0.16 & -0.21 & -0.17 & 0.03 & 0.27 \\
0.27 & 0.11 & -0.43 & 0.07 & 0.08 & -0.17 & 0.28 & -0.02 & -0.05 \\
0.27 & 0.11 & -0.43 & 0.07 & 0.08 & -0.17 & 0.28 & -0.02 & -0.05 \\
0.30 & -0.14 & 0.33 & 0.19 & 0.11 & 0.27 & 0.03 & -0.02 & -0.17 \\
0.21 & 0.27 & -0.18 & -0.03 & -0.54 & 0.08 & -0.47 & -0.04 & -0.58 \\
0.01 & 0.49 & 0.23 & 0.03 & 0.59 & -0.39 & -0.29 & 0.25 & -0.23 \\
0.04 & 0.62 & 0.22 & 0.00 & -0.07 & 0.11 & 0.16 & -0.68 & 0.23 \\
0.03 & 0.45 & 0.14 & -0.01 & -0.30 & 0.28 & 0.34 & 0.68 & 0.18
\end{array}$$

S_0 =

$$\begin{array}{ccccccccc}
3.34 & & & & & & & & \\
 & 2.54 & & & & & & & \\
 & & 2.35 & & & & & & \\
 & & & 1.64 & & & & & \\
 & & & & 1.50 & & & & \\
 & & & & & 1.31 & & & \\
 & & & & & & 0.85 & & \\
 & & & & & & & 0.56 & \\
 & & & & & & & & 0.36
\end{array}$$

D_0 =

$$\begin{array}{rrrrrrrrr}
0.20 & -0.06 & 0.11 & -0.95 & 0.05 & -0.08 & 0.18 & -0.01 & -0.06 \\
0.61 & 0.17 & -0.50 & -0.03 & -0.21 & -0.26 & -0.43 & 0.05 & 0.24 \\
0.46 & -0.03 & 0.21 & 0.04 & 0.38 & 0.72 & -0.24 & 0.01 & 0.02 \\
0.54 & -0.23 & 0.57 & 0.27 & -0.21 & -0.37 & 0.26 & -0.02 & -0.08 \\
0.28 & 0.11 & -0.51 & 0.15 & 0.33 & 0.03 & 0.67 & -0.06 & -0.26 \\
0.00 & 0.19 & 0.10 & 0.02 & 0.39 & -0.30 & -0.34 & 0.45 & -0.62 \\
0.01 & 0.44 & 0.19 & 0.02 & 0.35 & -0.21 & -0.15 & -0.76 & 0.02 \\
0.02 & 0.62 & 0.25 & 0.01 & 0.15 & 0.00 & 0.25 & 0.45 & 0.52 \\
0.08 & 0.53 & 0.08 & -0.03 & -0.60 & 0.36 & -0.04 & -0.07 & -0.45
\end{array}$$

The reader can verify that:

$X = T_0 S_0 D_0'$ (except for small rounding errors)
T_0 nas orthogonal, unit length columns so $T_0 T_0' = I$
D_0 has orthogonal, unit length columns so $D_0 D_0' = I$

We now approximate X keeping only the first two singular values and the corresponding columns from the T and D matrices. (Note that these are the T and D coordinates used to position the 12 terms and nine documents, respectively, in the two-dimensional representation of Figure 1.) In this *reduced model*,

$$X \approx \hat{X} = TSD'$$

$\hat{X}$ =

$$\begin{array}{rr|rr|rrrrrrrrr}
T & & S & & & & & & D & & & & \\
0.22 & -0.11 & 3.34 & & 0.20 & 0.61 & 0.46 & 0.54 & 0.28 & 0.00 & 0.02 & 0.02 & 0.08 \\
0.20 & -0.07 & & 2.54 & -0.06 & 0.17 & -0.13 & -0.23 & 0.11 & 0.19 & 0.44 & 0.62 & 0.53 \\
0.24 & 0.04 & & & & & & & & & & & \\
0.40 & 0.06 & & & & & & & & & & & \\
0.64 & -0.17 & & & & & & & & & & & \\
0.27 & 0.11 & & & & & & & & & & & \\
0.27 & 0.11 & & & & & & & & & & & \\
0.30 & -0.14 & & & & & & & & & & & \\
0.21 & 0.27 & & & & & & & & & & & \\
0.01 & 0.49 & & & & & & & & & & & \\
0.04 & 0.62 & & & & & & & & & & & \\
0.03 & 0.45 & & & & & & & & & & &
\end{array}$$

Multiplying out the matrices TSD' gives the following estimate of X, $\hat{X}$.

$\hat{X}$ =

$$\begin{array}{rrrrrrrrr}
0.15 & 0.40 & 0.38 & 0.47 & 0.18 & -0.05 & -0.12 & -0.16 & -0.09 \\
0.14 & 0.37 & 0.33 & 0.40 & 0.16 & -0.03 & -0.07 & -0.10 & -0.04 \\
0.15 & 0.51 & 0.36 & 0.41 & 0.24 & 0.02 & 0.06 & 0.09 & 0.12 \\
0.26 & 0.84 & 0.61 & 0.70 & 0.39 & 0.03 & 0.08 & 0.12 & 0.19 \\
0.45 & 1.23 & 1.05 & 1.27 & 0.56 & -0.07 & -0.15 & -0.21 & -0.05 \\
0.16 & 0.58 & 0.38 & 0.42 & 0.28 & 0.06 & 0.13 & 0.19 & 0.22 \\
0.16 & 0.58 & 0.38 & 0.42 & 0.28 & 0.06 & 0.13 & 0.19 & 0.22 \\
0.22 & 0.55 & 0.51 & 0.63 & 0.24 & -0.07 & -0.14 & -0.20 & -0.11 \\
0.10 & 0.53 & 0.23 & 0.21 & 0.27 & 0.14 & 0.31 & 0.44 & 0.42 \\
-0.06 & 0.23 & -0.14 & -0.27 & 0.14 & 0.24 & 0.55 & 0.77 & 0.66 \\
-0.06 & 0.34 & -0.15 & -0.30 & 0.20 & 0.31 & 0.69 & 0.98 & 0.85 \\
-0.04 & 0.25 & -0.10 & -0.21 & 0.15 & 0.22 & 0.50 & 0.71 & 0.62
\end{array}$$

There are two things to note about the $\hat{X}$ matrix. (1) It does not exactly match the original term by document matrix X (it gets closer and closer as more and more singular values are kept). (2) This is what we want; we don't want perfect fit since we think some of the 0's in X should be 1 and vice versa.

Acknowledgments

We would like to thank several colleagues who have been involved in various aspects of this project. Laura Beck has contributed extensively to programming, data collection, and analyses. Lynn Streeter, Karen Lochbaum, Bob Allen and Steve Hanson have all used our programs and provided useful feedback. We thank Mike Lesk for advice and comments on previous drafts of the article, and Ram Gnanadesikan, John Kettenring, and John Tukey for consultation on statistical questions.

References

Amsler, R. (1984). Machine-readable dictionaries. In *Annual Review of Information Science and Technology (ARIST)*, *19*, 161–209.

Atherton, P. & Borko, H. (1965). A test of factor-analytically derived automated classification methods. AIP Rept. AIP-DRP 65-1.

Baker, F. B. (1962). Information retrieval based on latent class analysis. *Journal of the ACM*, 9, 512–521.

Bates, M. J. (1986). Subject access in online catalogs: A design model. *JASIS*, *37*, 357–376.

Borko, H. & Bernick, M. D. (1963). Automatic document classification. *Journal of the ACM*, *10*, 151–162.

Carroll, J. D. & Arabie, P. (1980). Multidimensional scaling. In M. R. Rosenzweig and L. W. Porter (Eds.). *Annual Review of Psychology*, *31*, 607–649.

Carroll, J. D. & Chang, J. J. (1970). Analysis of individual differences in multidimensional scaling via an N-way generalization of "Eckart-Young" decomposition. *Psychometrika*, *35*, 283–319.

Choueka, Y. & Lusignan, S. (1985). Disambiguation by short contexts. *Computers and the Humanities*, *19*, 147–157.

Coombs, C. H. (1964). *A theory of data*. New York: Wiley.

Cullum, J., Willoughby, R. A., & Lake, M. (1983). A Lanczos algorithm for computing singular values and vectors of large matrices. *SIAM J. Sci. Stat. Comput.*, *4*, 197–215.

Desarbo, W. S. & Carroll, J. D. (1985). Three-way metric unfolding via alternating weighted least squares. *Psychometrika*, *50*, 275–300.

Fidel, R. (1985). Individual variability in online searching behavior. In C. A. Parkhurst (Ed.), *ASIS'85: Proceedings of the ASIS 48th Annual Meeting*, Vol. 22, 69–72.

Forsythe, G. E., Malcolm, M. A., & Moler, C. B. (1977). *Computer methods for mathematical computations* (Chapter 9: Least squares and the singular value decomposition). Englewood Cliffs, NJ: Prentice Hall.

Furnas, G. W. (1980). Objects and their features: The metric representation of two-class data. Ph.D. Dissertation, Stanford University.

Furnas, G. W. (1985). Experience with an adaptive indexing scheme. In *Human Factors in Computers Systems, CHI'85 Proceedings*, 130–135.

Furnas, G. W., Landauer, T. K., Gomez, L. M., & Dumais, S. T. (1987). The vocabulary problem in human-system communications. *Communications of the ACM, 30*, 964–971.

Golub, G. H., Luk, F. T., & Overton, M. L. (1981). A block Lanczos method for computing the singular values and corresponding singular vectors of a matrix. *ACM Transactions on Mathematical Software, 7*, 149–169.

Gomez, L. M., Lochbaum, C. C., & Landauer, T. K. (in press). All the right words: Finding what you want as a function of indexing vocabulary. *JASIS*, in press.

Harshman, R. A. (1970). Foundations of the PARAFAC procedure: Models and conditions for an "explanatory" multi-modal factor analysis. *UCLA Working Phonetics Paper*.

Harshman, R. A. & Lundy, M. E. (1984a). The PARAFAC model for three-way factor analysis and multidimensional scaling. In H. G. Law, C. W. Snyder, Jr., J. A. Hattie, and R. P. McDonald (Eds.), *Research methods for multimode data analysis*. New York: Praeger.

Harshman, R. A. & Lundy, M. E. (1984b). Data preprocessing and the extended PARAFAC model. In H. G. Law, C. W. Snyder, Jr., J. A. Hattie, and R. P. McDonald (Eds.), *Research methods for multimode data analysis*. New York: Praeger.

Heiser, W. J. (1981). *Unfolding analysis of proximity data*. Leiden, The Netherlands: Reprodienst Psychologie RUL.

Jardin, N. & van Rijsbergen, C. J. (1971). The use of hierarchic clustering in information retrieval. *Information Storage and Retrieval, 7*, 217–240.

Jones, W. P. & Furnas, G. W. (1987). Pictures of relevance. *JASIS, 38*, 420–442.

Koll, M. (1979). An approach to concept-based information retrieval. *ACM SIGIR Forum, 13*, 32–50.

Kruskal, J. B. (1978). Factor analysis and principal components: Bilinear methods. In H. Kruskal and J. M. Tanur (Eds.), *International encyclopedia of statistics*. New York: Free Press.

Lesk, M. E. (1986). How to tell a pine cone from an ice cream cone. In *Proceedings of ACM SIGDOC Conference*, 24–26.

Liley, O. (1954). Evaluation of the subject catalog. *American Documentation, 5*, 41–60.

Ossorio, P. G. (1966). Classification space: A multivariate procedure for automatic document indexing and retrieval. *Multivariate Behavior Research*, 479–524.

Raghavan, V. & Wong, S. (1986). A critical analysis of vector space model for information retrieval. *JASIS, 37*, 279–288.

Salton, G. (1968). *Automatic information organization and retrieval*. New York: McGraw-Hill.

Salton, G. & McGill, M. J. (1983). *Introduction to modern information retrieval*. New York: McGraw-Hill.

Sparck Jones, K. (1971). *Automatic keyword classification for information retrieval*. London: Buttersworth.

Sparck Jones, K. (1972). A statistical interpretation of term specificity and its applications in retrieval. *Journal of Documentation, 28*, 11–21.

Tarr, D. & Borko, H. (1974). Factors influencing inter-indexer consistency. In *Proceedings of the ASIS 37th Annual Meeting*, Vol. *11*, 50–55.

van Rijsbergen, C. J. (1977). A theoretical basis for the use of co-occurrence data in information retrieval. *Journal of Documentation, 33*, 106–119.

Voorhees, E. (1985). The cluster hypothesis revisited. *Proceedings of SIGIR*, 188–196.

UNITED STATES
NUCLEAR REGULATORY COMMISSION
REGION I
475 ALLENDALE ROAD
KING OF PRUSSIA, PENNSYLVANIA 19406

Docket No. 030-30348 SEP 18 1990 License No. 29-28137-01

Marc Associates, Inc
ATTN: Arthur Bage, President
Route 72, PO Box 479
Chatsworth, NJ 08019

Gentlemen:

Subject: Inquiry No. 90-001

This refers to a telephone inquiry by Mary Cahill of this office with Cheryl Jones of your staff on September 18, 1990. This inquiry concerned activities authorized by the above listed NRC license.

From this discussion, we understand the following:

☒ You have never possessed material authorized by this license and do not plan to acquire such material in the near future. We further understand that you will notify this office by telephone or in writing prior to acquiring licensed material.

□ You have never possessed material authorized by this license, but you plan to acquire such material in the near future.

□ You plan to send a letter to this office request , termination of your license. Please include the enclosed Certificate of Disposition of Materials with your letter.

If our understanding is incorrect, please inform us in writing.

In accordance with Section 2.790 of the NRC's "Rules of Practice", Part 2, Title 10, Code of Federal Regulations, a copy of this letter will be placed in the Public Document Room. No reply to this letter is required; however, we will be pleased to discuss any questions with you.

Your cooperation with us is appreciated.

Sincerely,

Mary A Cahill

Health Physicist
Nuclear Materials Safety Branch
Division of Radiation Safety
and Safeguards

r/o

Enclosure: □ Certificate of Disposition of Materials (Form NRC 314)

cc:
Region I Docket Room
State of New Jersey

OFFICIAL RECORD COPY

RETURN ORIGINAL TO
REGION I

IE:07