Design Criteria for Automatic Information Systems*

M. E. Lesk+ and G. Salton‡

1. Introduction

Considerable attention has been paid in recent years to the organization of information centers. Various plans have been advanced for the establishment of partly mechanized information and library centers, and recommendations have been drawn up specifying the organization of a national document handling system. [1,2] In general, such plans stipulate use of a given equipment complex to store the information to be searched. Provision is normally made for introducing search requests from a variety of input stations, some of which may be situated far away from the central equipment, and users are often allowed to submit their requests asynchronously, and independently of each other.

Nearly all of those proposals are, moreover, based on a number of underlying assumptions, which though unproved and unaccompanied by supporting evidence, are nevertheless stated with great forcefulness and considered to be axiomatic. The principal assumptions may be stated as follows:

a) a computer cannot perform the intellectual work required to analyze the content of a document, and information centers must therefore rely on a large staff of human subject experts to assign keywords to all items stored in the system;

---

* This study was supported in part by the National Science Foundation under grants GN-360 and GN-495.

+ Aiken Computation Laboratory, Harvard University, Cambridge, Mass., 02138.

‡ Dept. of Computer Science, Cornell University, Ithaca, New York, 14850.

b) the intellectual aids to be used as part of the manual analysis and indexing procedure, including dictionaries, thesauruses, and hierarchical subject arrangements are best prepared and maintained by committees of experts in the subject areas under consideration;

c) the users of the service, being unaware of system restrictions and operations, should not submit search requests directly to the system but must work through human intermediaries who analyze the query statement and prepare suitable search formulations for introduction into the program.

A system organization based on these principles leads to a service in which only the search operations themselves are mechanized (that is, the comparisons between analyzed information items and analyzed search requests), but most other operations are carried out semimanually or manually. It also results in an information system which suffers from so many built-in weaknesses that adequate service to the users cannot ever be expected.

The first weakness is the well-known scarcity and increasing unavailability of subject experts who are willing and able to perform a manual content analysis of the documents and search requests. This simple fact results in a continuing crisis atmosphere in existing nonconventional search systems, a situation which may be expected to grow more severe as time progresses. The second weakness is the inadequacy of the presently available dictionaries and authority lists which are used to control the assignment of subject identifiers to the stored information. These dictionaries are often produced as a result of so many compromises among various expert committees, that the final product reflects no consistent point of view, and is difficult to utilize effectively. The third weakness is the absence of meaningful user interaction with the system, so that individual

user needs and reactions by users to initial search efforts cannot usefully be taken into account in order to improve the service.

The SMART document retrieval system which has been operating on an IBM 7094 for the last two years has been used extensively to test a large variety of automatic retrieval procedures, including fully automatic information analysis methods, automatic procedures for dictionary construction, and iterative search techniques based on user interaction with the system.[3,4,5,6] The evaluation results indicate that presently held assumptions concerning the design of information systems are untenable, and point the way to alternative design criteria. Some of the experiments conducted with the SMART system are outlined briefly, and the principal results are described in the remainder of this study.

2. The SMART Experiments

SMART is a fully automatic document retrieval system operating on the IBM 7094. The system does not rely on manually assigned keywords or index terms for the identification of documents and search requests, nor does it use primarily the frequency of occurrence of certain words or phrases included in the document texts. Instead, the system goes beyond simple word-matching procedures by using a variety of intellectual aids in the form of synonym dictionaries, hierarchical arrangements of subject identifiers, statistical and syntactic phrase generating methods, and the like, in order to obtain the content identifications useful for the retrieval process.

Stored documents and search requests are then processed <u>without any prior manual analysis</u> by one of several hundred automatic content analysis methods, and those documents which most nearly match a given search request

are identified. Specifically, a correlation coefficient is computed to indicate the degree of similarity between each document and each search request, and documents are then ranked in decreasing order of the correlation coefficient.[3,4,5] A typical search request processed by the system is shown in Fig. 1. Three analyzed forms of this request, produced respectively by a word stem identification process (null thesaurus), a synonym dictionary look-up (regular thesaurus), and a phrase identification method (statistical phrases), are shown in Fig. 2. Finally a typical output product listing documents in decreasing correlation order with the request is shown in Fig. 3.

The system may be controlled by the user in that a search request can be processed first in a standard mode. The user can then analyze the output obtained and depending on the information returned to the system as a result of previous search operations, the request can be reprocessed under altered conditions. The new output can again be examined, and the search can be interated until the right kind and amount of information are obtained.[6,7]

The SMART systems organization makes it possible to evaluate the effectiveness of the various processing methods by comparing the output obtained from a variety of different runs. This is achieved by processing the same search requests against the same document collections several times, while making selected changes in the analysis procedures between runs. By comparing the performance of the search requests under different processing conditions, it is then possible to determine the relative effectiveness of the various analysis methods.

The actual evaluation calculations are based on the standard recall

PAGE 345

ENGLISH TEXT PROVIDED FOR DOCUMENT DIFFERNTL EQ SEPT. 28, 1964

GIVE ALGORITHMS USEFUL FOR THE NUMERICAL SOLUTION 1
OF ORDINARY DIFFERENTIAL EQUATIONS AND PARTIAL DIFFER- 1
ENTIAL EQUATIONS ON DIGITAL COMPUTERS. EVALUATE THE 1
VARIOUS INTEGRATION PROCEDURES (TRY RUNGE-KUTTA, 2
MILNE-S METHOD) WITH RESPECT TO ACCURACY, STABILITY, 2
AND SPEED. 2

TYPICAL SEARCH REQUEST

Fig. 1

OCCURRENCES OF CONCEPTS AND PHRASES IN DOCUMENTS — SEPTEMBER 28, 1964

DOCUMENT — CONCEPT, OCCURS — PAGE 17

DOCUMENT	CONCEPT, OCCURS														
DIFFERNTL EQ	ACCUR	12	ALGORI	12	COMPUT	12	DIFFER	24	DIGIT	12					NULL
	EQU	24	EVALU	12	GIVE	12	INTEGR	12	METHOD	12					
	NUMER	12	ORDIN	12	PARTI	12	PROCED	12	RUNGE-	12					THESAURUS
	SOLUT	12	SPEED	12	STABIL	12	USE	12	VARIE	12					
DIFFERNTL EQ	4EXACT	12	8ALGOR	12	13CALC	18	71EVAL	6	92DIGI	12					REGULAR
	110AUT	12	143UTI	12	176SOL	12	179STD	12	181QUA	24					
	269ELI	4	274DIF	36	356VEL	12	357YAW	4	384TEG	12					THESAURUS
	428STB	4	505APP	24											
DIFFERNTL EQ	4EXACT	12	8ALGOR	12	13CALC	18	71EVAL	6	92DIGT	12					STATISTICAL
	110AUT	12	143UTI	12	176SOL	12	179STD	12	181QUA	24					PHRASES
	269ELI	4	274DIF	36	356VEL	12	357YAW	4	375NUM	36					LOOK-UP
	379DIF	72	384TEG	12	428STB	4	505APP	24							

INDEXING PRODUCTS FOR "DIFFERENTIAL EQUATIONS"

Fig. 2

CURRENT REQUEST - *LIST DIFFERNTL EQ NUMERICAL DIGITAL SOLN OF DIFFERENTIAL EQUATIONS

REQUEST *LIST DIFFERNTL EQ NUMERICAL DIGITAL SOLN OF DIFFERENTIAL EQUATIONS

GIVE ALGORITHMS USEFUL FOR THE NUMERICAL SOLUTION OF ORDINARY DIFFERENTIAL EQUATIONS AND PARTIAL DIFFERENTIAL EQUATIONS ON DIGITAL COMPUTERS . EVALUATE THE VARIOUS INTEGRATION PROCEDURES (E.G. RUNGE--KUTTA, MILNE-S METHOD) WITH RESPECT TO ACCURACY, STABILITY, AND SPEED .

ANSWER	CORRELATION	IDENTIFICATION
384STABILITY	0.6675	STABILITY OF NUMERICAL SOLUTION OF DIFFERENTIAL EQUATIONS W. E. MILNE AND R. R. REYNOLDS (OREGON STATE COLLEGE) J. ASSOC. FOR COMPUTING MACH. VOL 6 PP 196-203 (APRIL, 1959)

ANSWER	CORRELATION	IDENTIFICATION
360SIMULATIN	0.5758	SIMULATING SECOND-ORDER EQUATIONS D. G. CHADWICK (UTAH STATE UNIV.) ELECTRONICS VOL 32 P 64 (MARCH 6, 1959)

ANSWER	CORRELATION	IDENTIFICATION
200SOLUTION	0.5663	SOLUTION OF ALGEBRAIC AND TRANSCENDENTAL EQUATIONS ON AN AUTOMATIC DIGITAL COMPUTER G.N. LANCE (UNIV. OF SOUTHAMPTON) J. ASSOC. FOR COMPUTING MACH., VOL 6, PP 97-101, JAN., 1959

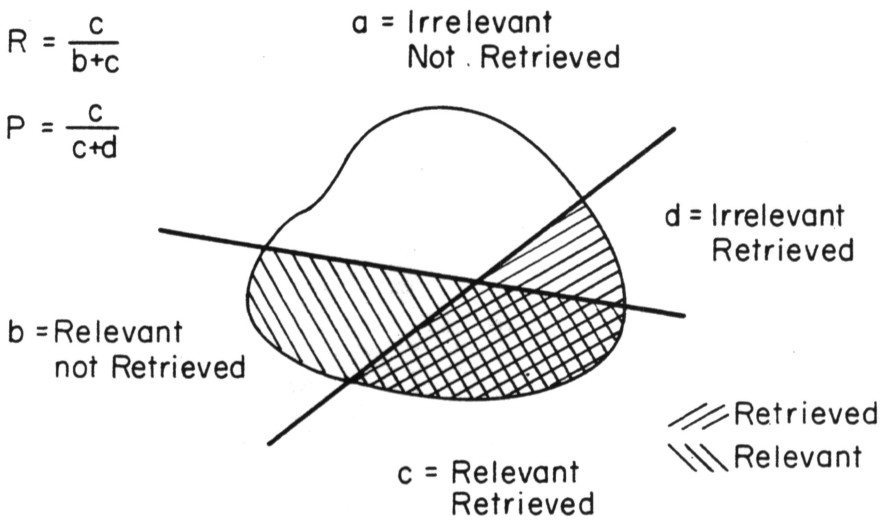
EXCERPT FROM LONG FORM OUTPUT (REGULAR THESAURUS)

Fig. 3

and precision measures, where the recall is defined as the proportion of relevant matter retrieved, while precision is the proportion of retrieved material actually relevant. If a dual cut is made through the document collection to distinguish retrieved items from nonretrieved on the one hand, and relevant items from nonrelevant ones on the other, the two measures may be defined as shown in Fig. 4. The computation of these measures is straightforward only if exhaustive relevance judgments are available for each document with respect to each search request, and if the cut-off value distinguishing retrieved from nonretrieved material can be unambiguously determined.[8,9,10]

In the evaluation work carried out with the SMART system, manually derived, exhaustive relevance judgments could be used since the document collections processed are all relatively small. Moreover, the choice of a unique cut-off could be avoided by computing the precision for various recall values, and exhibiting a plot showing recall against precision. Recall-precision graphs, such as those shown in the remainder of this study, have been criticized for a variety of reasons,[11] but they are very effective to summarize the performance of retrieval methods averaged over many search requests, and they can be used advantageously to select analysis methods which fit certain specific operating ranges. Thus, if it is desired to pick a procedure which favors the retrieval of all relevant material, then one must concentrate on the high recall region; similarly, if only relevant material is wanted, the high precision region is of importance. In general, it is possible to obtain high recall only at a substantial cost in precision, and vice-versa.[8,9,10]

The following document collections have been used in the experiments



PARTITIONING OF DOCUMENT COLLECTION

Fig. 4

with the SMART system:

a) IRE - 1 : a set of about 400 abstracts of documents in the computer literature published in 1959, used with approximately 20 search requests;

b) IRE - 2 : a set of about 400 abstracts of documents in the computer literature published in 1960 and 1961, used with approximately 20 search requests;

c) ADI : a set of 82 short papers in documentation, each approximately 2000 words long, presented at the 1963 Annual Meeting of the American Documentation Institute, and processed against 35 search requests;

d) Cranfield - 1 : a set of 200 abstracts of documents in aeronautical engineering previously used by the Aslib-Cranfield project [12], and processed against 42 search requests;

e) Cranfield - 2 : a set of 1200 additional document abstracts in aeronautical engineering, similar to the abstracts included in the preceding collection.

It is seen that these collections fall into three distinct subject areas: computer science, documentation, and aeronautical engineering. The ADI collection in documentation is of particular interest because full papers are available rather than only document abstracts. The Cranfield collections, on the other hand, are the only ones which are also manually indexed by subject experts, thus permitting a comparison of the standard keyword search procedures with the automatic text processing methods.

The evaluation results obtained with the first four of these collections are summarized in the next section.

3. Evaluation Results and Design Criteria

In attempting to generate useful criteria for the design of information systems, a number of obvious questions suggest themselves: first, can automatic text processing methods be used effectively to replace a manual content analysis; if so, what part or parts of a document should be incorporated in the automatic procedure; is it necessary to provide vocabulary normalization methods to eliminate ambiguities caused by homographs and synonymous word groups; should such a normalization be handled by means of a specially constructed dictionary, or is it possible to replace thesauruses completely by statistical word association methods; what dictionaries can most effectively be used for vocabulary normalization; is it important to provide hierarchical arrangements of subject categories as is done in many library classification systems; what should be the role of the user in formulating and controlling the search procedure. These and many other questions are considered in the evaluation process described in the remainder of this section.

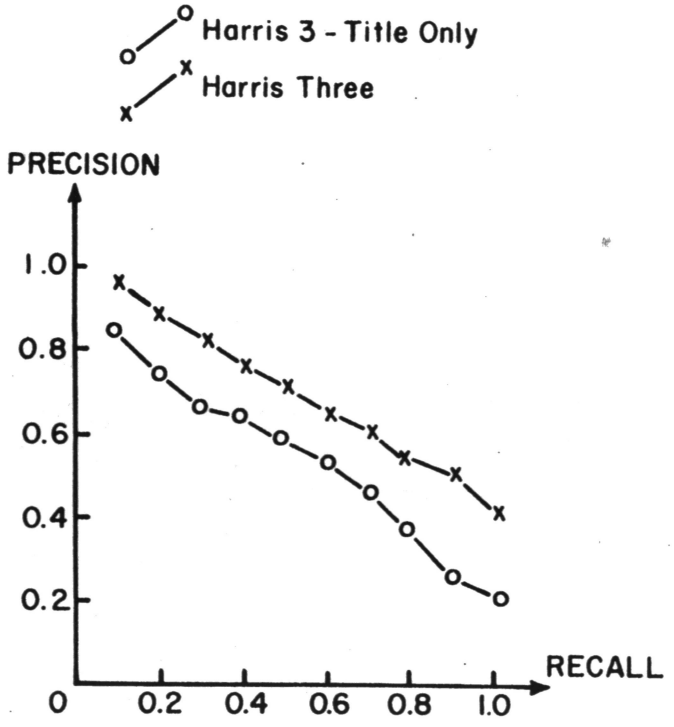
A) Indexing Depth and Document Length

In a manual system, where each information item is identified by a few carefully chosen keywords, the presence or absence of a given keyword becomes of crucial importance, since failure to provide a certain needed keyword may mean the difference between a retrievable item and one which is not. In an automatic text processing system, it is possible to generate for each item many different information identifiers, as seen in Fig. 2 for the request of Fig. 1; the importance of each individual identifier is then much reduced since a small number of poorly chosen terms are often offset by the much larger number of correct ones.

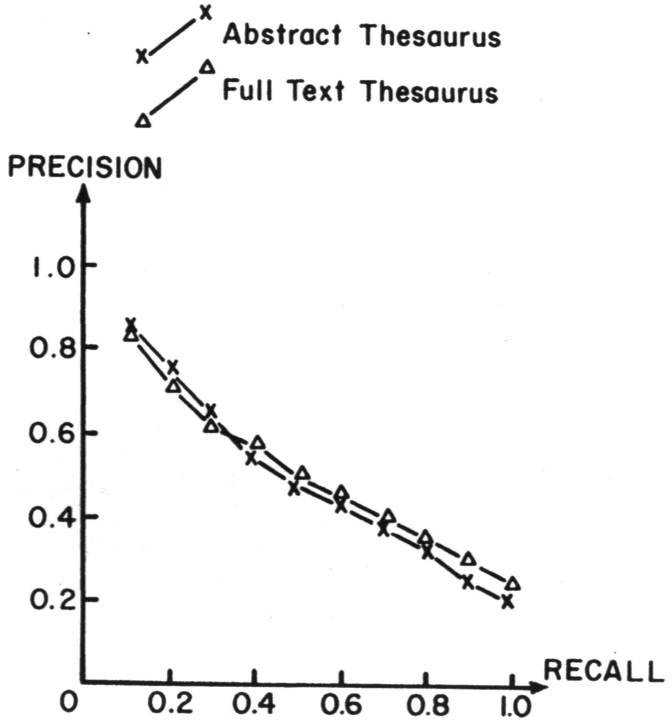
A second principal difference between manual and automatic information analysis systems is the relative difficulty in manual systems of discriminating among keywords by weights assigned to reflect their relative importance. This results in the "all or nothing" situation where a given identifier is either present or not, and each identifier is considered to be equally important. In an automatic system, on the other hand, it is easy to assign weights to individual identifiers, as shown in Fig. 2. These weights can be derived in part by using the frequency of occurrence of the original text words, and in part as a function of the various dictionary mapping procedures. Thus, ambiguous terms which in a synonym dictionary correspond to many different concept classes, can be weighted less than unambiguous terms.

The relative usefulness of analyzing document sections of varying lengths, and of utilizing weighted terms is reflected in the output of Figs. 5 and 6. These recall-precision graphs exhibit output averaged over 17 search requests for the IRE - 2 collection and over 35 requests for the ADI material. Since it is in general desirable to get both high recall (that is, to retrieve most of what is relevant) and high precision (that is, to retrieve very little that is irrelevant), the region of importance is the upper right-hand corner of each graph. The more effective a given retrieval algorithm, the smaller will be the distance between the corresponding recall-precision curve and the 1:1 recall-precision point.

Fig. 5(a) shows a comparison of a "title only" option, where only the titles of documents are used in the analysis with a "full abstract" option. In both cases, the word stems originally extracted from document titles and document abstracts were first looked-up in a synonym dictionary



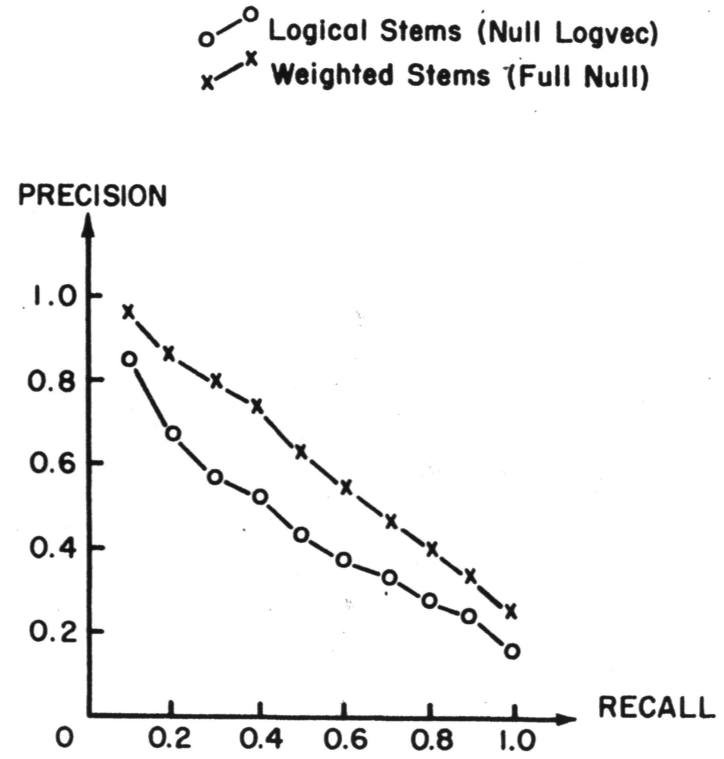
a) Title versus Abstract Process (IRE - 2)



b) Abstract versus Full Text (ADI collection)

Comparison Based on Document Length

Fig. 5



Effectiveness of Weighted Word Stems
(IRE-2 collection 17 requests)

Fig. 6

(called "Harris 3" in Fig. 5), to provide vocabulary normalization before the actual word matching operation. The curve of Fig. 5(a) makes clear how superior the full abstract process is compared with the title procedure. If the text words had been matched directly, without a thesaurus intermediary, the discrepancy between the two procedures would be even larger.

The output of Fig. 5(b) shows that a further improvement is obtainable if full text is used, rather than only abstracts, particularly for the high recall region. However, the improvement is much smaller here, and in actual practice it would seem that the additional problems arising from a full text process can be avoided by restricting the procedure to abstracts and summaries, unless a clear requirement exists for a high recall performance.

The output of Fig. 5 then leads to the following rule:

Rule 1 : The use of document titles alone for purposes of information analysis results in poor retrieval performance compared with the use of abstracts or full text.

Rule 1 is of particular interest because of the widespread advocacy of permuted title indexes (also known as KWIC indexes) for information search and retrieval purposes.

Fig. 6 shows the improvement obtainable by using weighted word stems, compared with unweighted stems. It is clear from the figure that term weights are essential for retrieval purposes, and it can be inferred that one of the main drawbacks of presently operating keyword search systems is the lack of discrimination between terms of varying importance. Rule 2 can then be stated as follows:

Rule 2 : The use of information identifiers which are weighted in accordance with their presumed importance leads to large-scale improvements in retrieval effectiveness, compared with the use of unweighted terms.

B) Synonym Recognition

One of the perennial problems in automatic language analysis is the question of language variability among authors, and the linguistic ambiguities which result. A large number of experiments have therefore been performed using a variety of synonym dictionaries for each of the three subject fields under study ("Harris 2" and "Harris 3" dictionaries for the computer literature, "Quasi-synonym" or "QS" lists for aeronautical engineering, and regular thesaurus for documentation). An excerpt of such a synonym dictionary for the computer literature is shown in Fig. 7 for the concept class numbers 408 to 416. Use of such a synonym dictionary permits the replacement of a variety of related terms by the corresponding concept classes, thus ensuring the retrieval of documents dealing with the "manufacture of transistor diodes" when the query deals with the "production of solid state rectifiers".

The output of Fig. 8 shows that considerable improvements in performance are obtainable by means of suitably constructed synonym dictionaries. The improvement is smallest for the Cranfield collection because the dictionary available for this collection was not originally constructed for retrieval purposes. This observation suggests that not all dictionaries are equally useful. Experiments conducted with the SMART system lead to the following principles of dictionary construction [13]:

408 DISLOCATION
JUNCTION
MINORITY-CARRIER
N-P-N
P-N-P
POINT-CONTACT
RECOMBINE
TRANSITION
UNIJUNCTION

409 BLAST-COOLED
HEAT-FLOW
HEAT-TRANSFER

410 ANNEAL
STRAIN

411 COERCIVE
DEMAGNETIZE
FLUX-LEAKAGE
HYSTERESIS
INDUCT
INSENSITIVE
MAGNETORESISTANCE
SQUARE-LOOP
THRESHOLD

412 LONGITUDINAL
TRANSVERSE

413 CAPACITANCE
IMPEDANCE-MATCHING
IMPEDANCE
INDUCTANCE
MUTUAL-IMPEDANCE
MUTUAL-INDUCTANCE
MUTUAL
NEGATIVE-RESISTANCE
POSITIVE-GAP
REACTANCE
RESIST
SELF-IMPEDANCE
SELF-INDUCTANCE
SELF

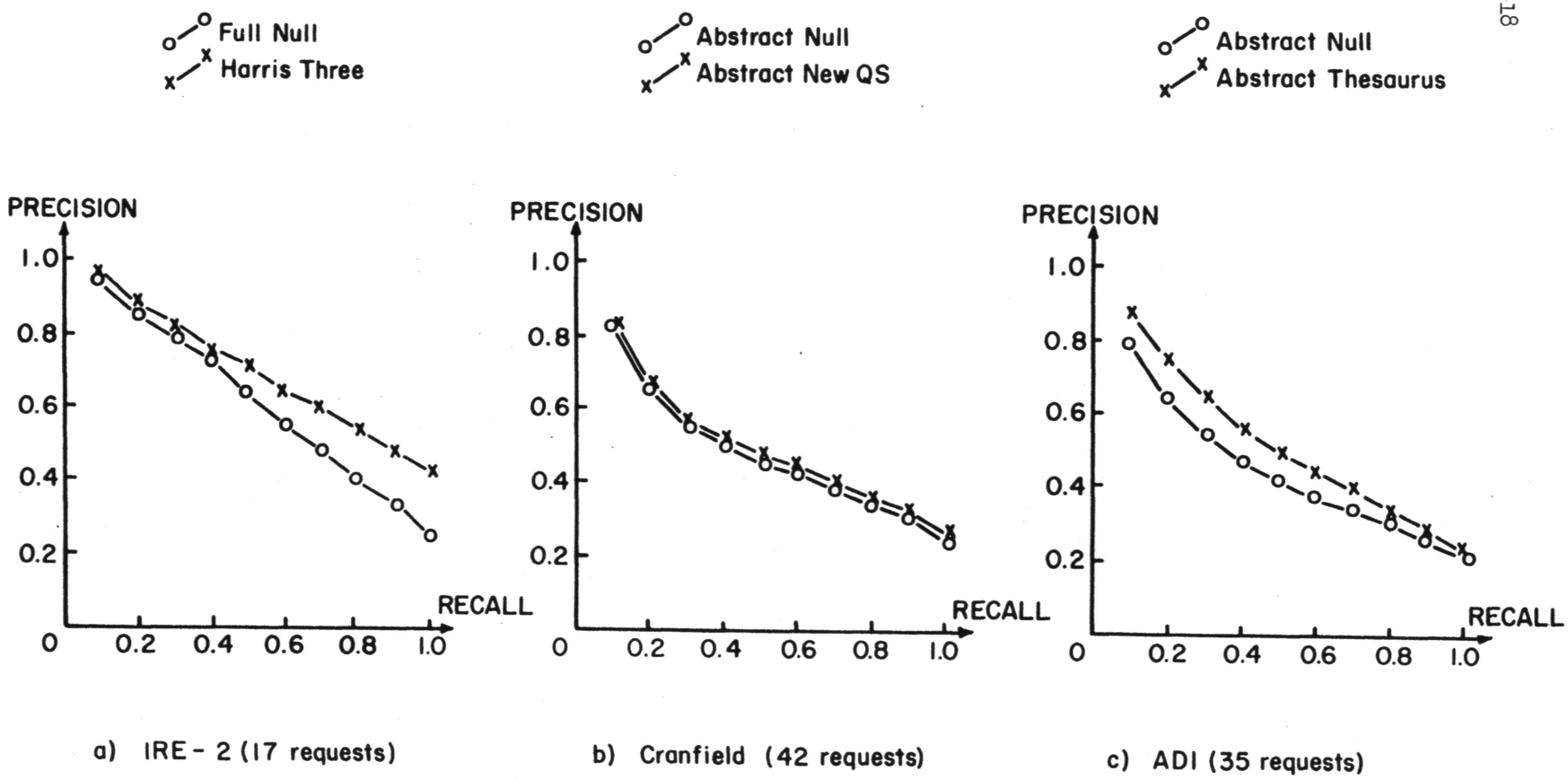
414 ANTENNA
KLYSTRON
PULSES-PER-BEAM
RECEIVER
SIGNAL-TO-RECEIVER
TRANSMITTER
WAVEGUIDE

415 CRYOGENIC
CRYOTRON
PERSISTENT-CURRENT
SUPERCONDUCT
SUPER-CONDUCT

416 RELAY

THESAURUS EXCERPT IN CONCEPT NUMBER ORDER

Fig. 7



Comparison of Synonym Recognition (Thesaurus) with Word Stem Matching Process

Fig. 8

a) very rare terms which occur in a representative sample document collection with insufficient frequency should not be included in the synonym dictionary, since such terms will not provide many matches between the stored items and the search requests;

b) very common high-frequency terms should either be eliminated, since they provide little discrimination, or shoule be placed into synonym classes of their own, so that they cannot submerge other terms which would be grouped with them;

c) terms which have no special significance in a given technical subject area (such as "begin", "indicate", "system", "automatic", etc.) should not be included;

d) ambiguous terms, such as for example "base", should be coded only for those senses which are likely to occur in the subject area being considered;

e) each group of synonymous terms should account for approximately the same total frequency of occurrence of the corresponding words in the document collection; this ensures that each identifier has approximately equal chance of being assigned to a given item.

These principles can be embodied in automatic programs for the construction of synonym dictionaries, using word frequency lists and concordances derived from a representative sample document collection.[13]

The experience gained with the various thesauruses constructed for the SMART system leads to Rule 3:

Rule 3 : Dictionaries providing synonym recognition are of considerable help in improving retrieval performance, particularly when they reflect the properties of the vocabulary under consideration.

C) Phrase Processing

The SMART system makes provision for the recognition of "phrases"

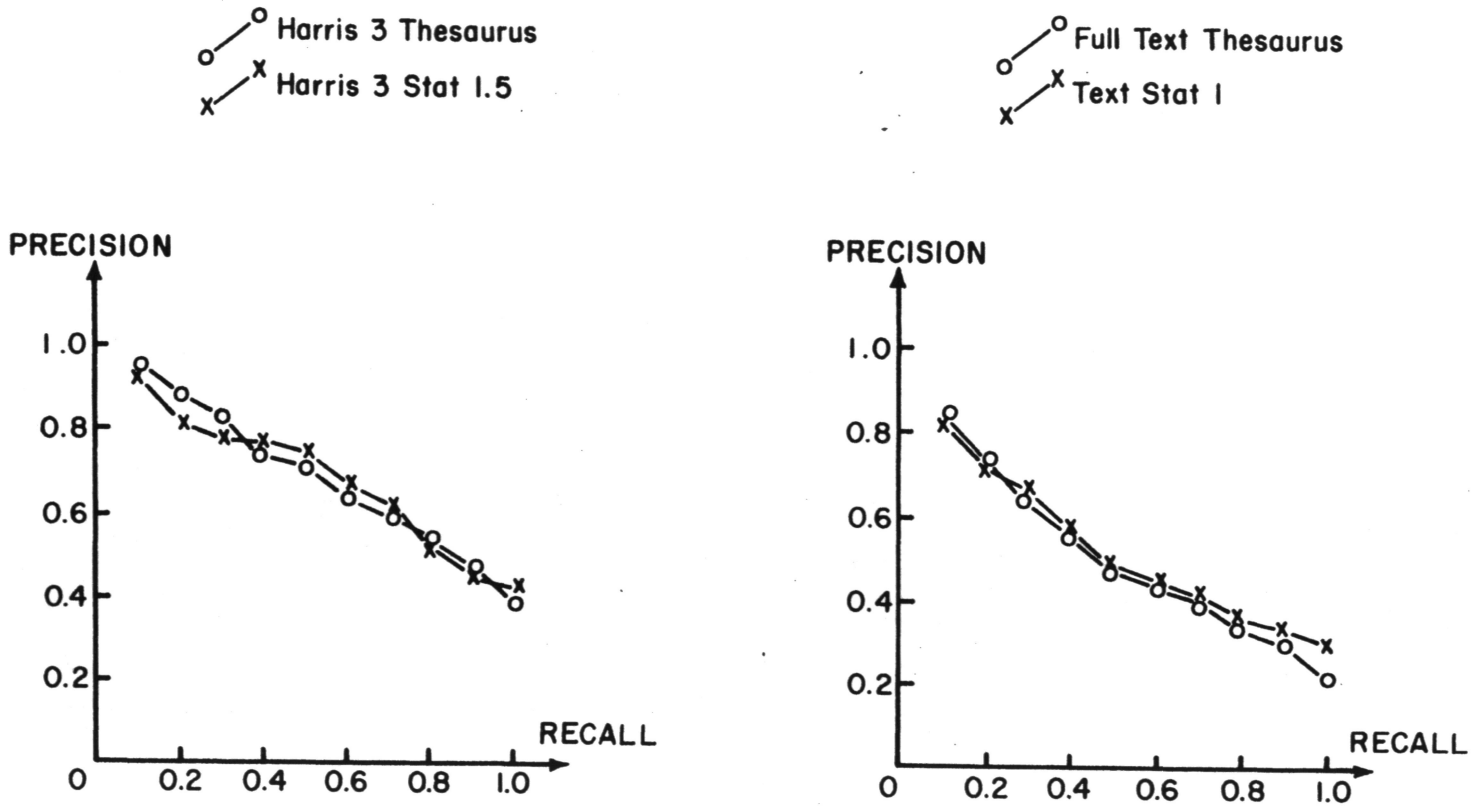
to identify documents and search requests, rather than only individual concepts alone. Thus if a given document contains the notion of "program" and the notion of "language", it might be tagged with the phrase "programming language". Phrases can be generated using a variety of strategies: for example, a phrase can be assigned any time the specified components co-occur in a given document, or in a given sentence of a document; alternatively, more restrictive phrase generation methods can be used by incorporating into the phrase generation process a syntactic recognition routine to check the syntactic compatibility between the phrase components before a phrase is actually accepted.[14]

In the SMART system, the normal phrase process uses a preconstructed dictionary of important phrases, and simple co-occurrence of phrase components, rather than syntactic criteria, are used to assign phrases to documents.* Phrases seem to be particularly useful as a means of incorporating into a document representation, terms whose individual components are not always meaningful by themselves. For example, "computer" and "control" are reasonably nonspecific, while "computer control" has a much more definite meaning in a computer science collection.

The output of Fig. 9 shows that phrases tend to improve recall at some expense in initial precision. This same effect was previously noted when the abstract processing was compared with full text in Fig. 5(b); it results from the fact that the simple process is good enough to retrieve the first few relevant documents (that is, in the high precision region), while the more sophisticated procedure is important if additional relevant documents are also wanted (that is, for high recall).

---

* Syntactic methods have, however, been used experimentally and sample results are published elsewhere.[6]



Comparison of Simple Thesaurus Process with Phrase Processing

Fig. 9

A phrase generation process which does not use a complete syntactic analysis of the phrase components may be expected to lead to many "false phrases", where components are combined which do not belong together (such as "information retrieval" in the sentence "for people in need of information retrieval is imperative"). The experimental evidence, reflected in the relatively poor performance of the syntactic process, makes it appear that such occurrences are very rare. This leads to Rules 4 and 5:

Rule 4 : Absolute accuracy in the analysis of every single item is not so important as the accumulation of a maximum number of correctly analyzed items. If a choice exists between a method which can produce one guaranteed correct content indication (syntactic analysis), and another which produces five indicators of which four are probably correct (statistical phrase process), the second is generally to be preferred.

Rule 5 : Simple phrase generation methods lead to a definite improvement in recall at the expense of some initial loss in precision in the low recall region.

D) Statistical Association Methods

Statistical association methods are those which use the co-occurrence frequency of two words, or two dictionary concepts, within a given document collection as an indication of a relationship between them.[15,16] Thus, if two given terms co-occur in many of the documents of a collection, or in many sentences within a given document, a non-zero correlation coefficient can be computed as a function of the number of co-occurrences. If this coefficient is sufficiently high, the two terms can be grouped, and can be assigned jointly to documents and search requests. Associative methods are

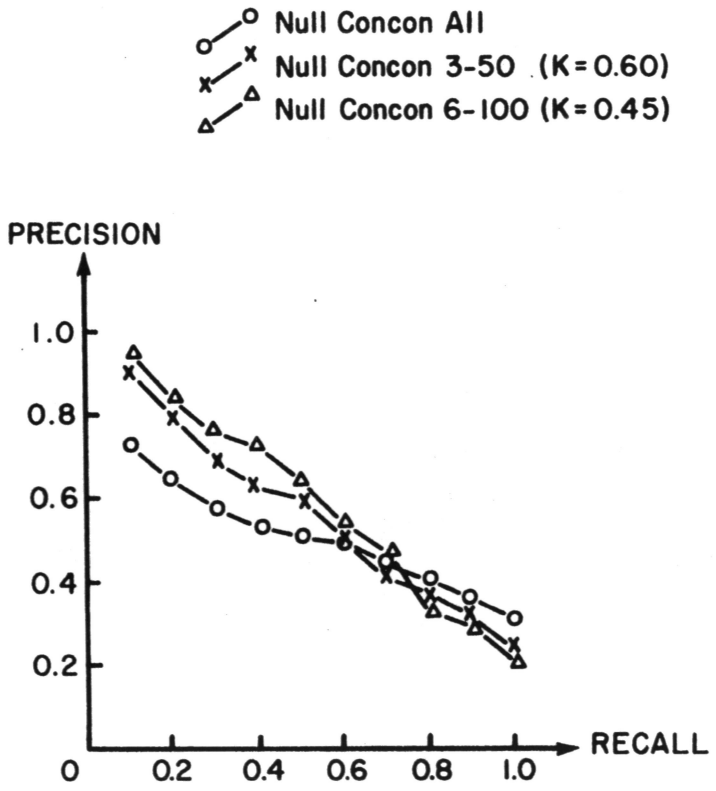
therefore comparable to thesaurus procedures, except that the word associations reflect strictly the vocabulary statistics of a given collection, whereas a thesaurus grouping may be expected to have a more general validity.

Many possible procedures exist for the generation of statistical word associations, leading to the identification of varying numbers of associated term pairs. Two main parameters are the cut-off value K in the association coefficient below which a statistical association is not recognized, and the frequency of occurrence of the terms being correlated. When all terms are correlated, no matter how low their frequency in the document collection, a great many spurious associations may be found; on the other hand, some correct associations will not be observable under any stricter conditions. The spurious associations result initially in low precision, but the few important associations will eventually produce improved recall in the high recall region. This is reflected in the curve for the "null concon all" process (concept-concept associations performed for all word stems regardless of frequency) of Fig. 10.

Increasingly more restrictive association procedures, applied first only to concepts in the frequency range 3 to 50, and then in the frequency range 6 to 100 eliminate many spurious associations, but also some correct ones. This results in a smaller initial loss in precision, but also in a poorer recall performance for high values. The output of Fig. 10 then confirms the following general rule:

Rule 6 : Deep indexing procedures which supply new information identifiers of which some are useful but many are not usually improve recall but depress precision.

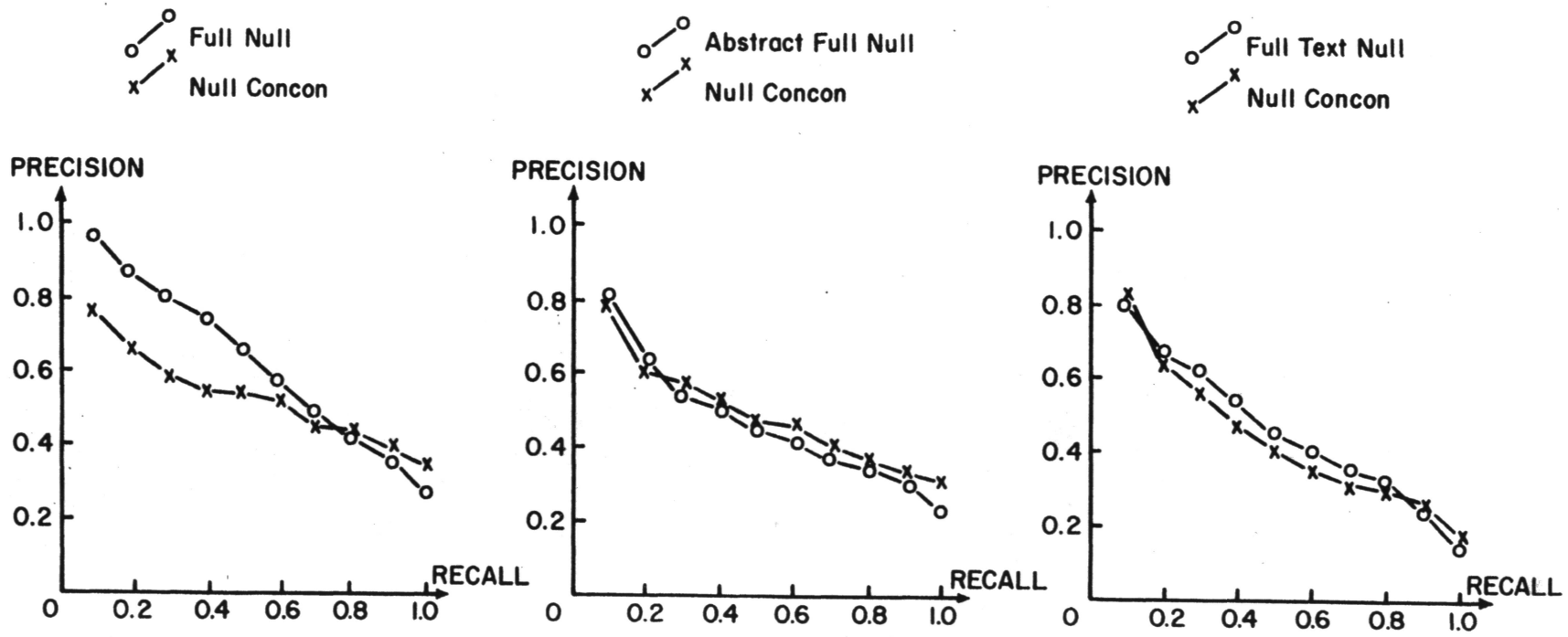
Fig. 11 exhibits the comparison between word-word association procedures



Comparison of Word-Word Association Strategies

(IRE-2 averaged over 17 requests)

Fig. 10



Comparison of Simple Word Stem Match (Full Null) with Statistical Word-Word Associations (Null Concon)

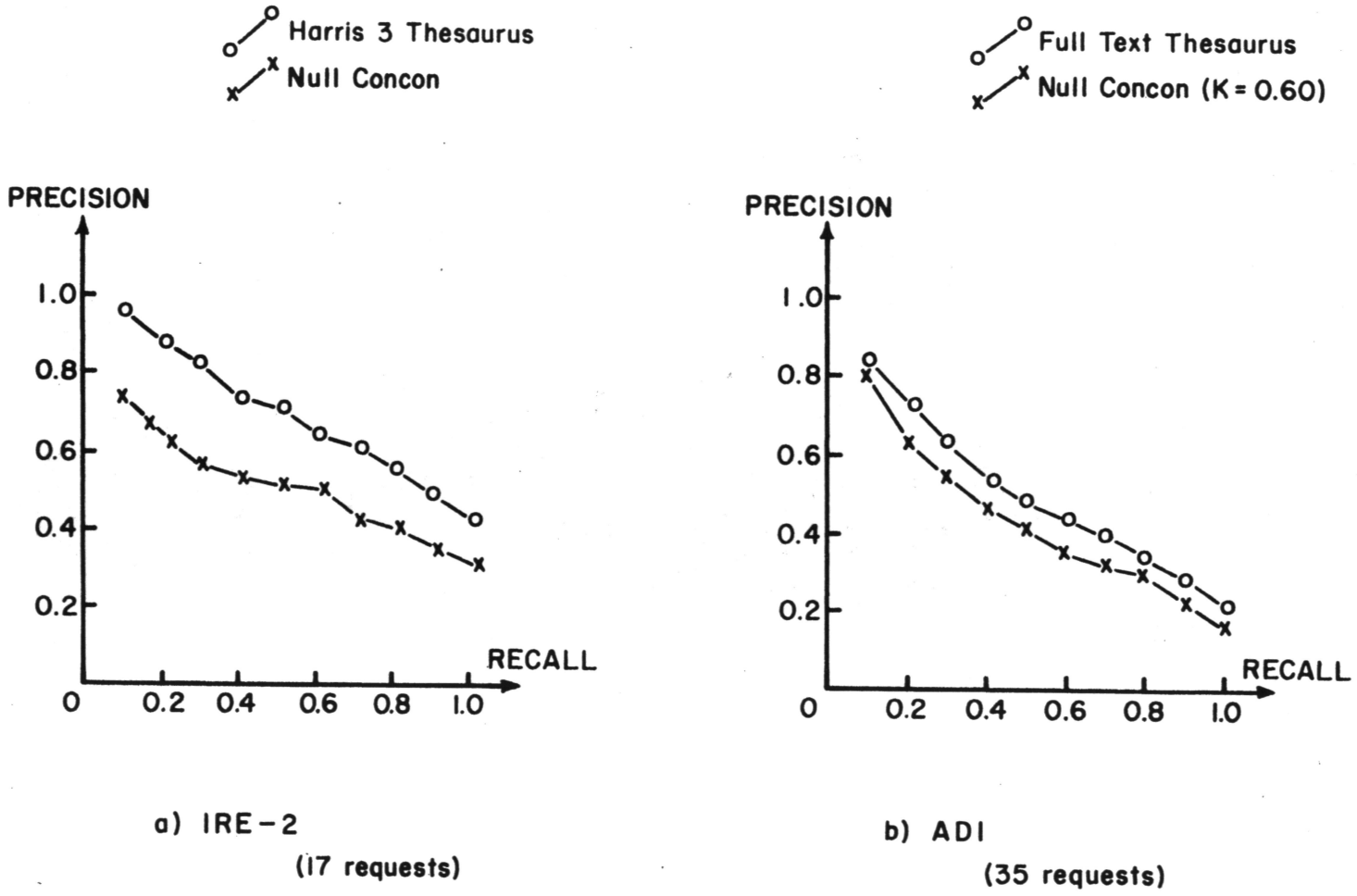
Fig. 11

(null concon), where associated word stems are added to the original stems available for content identification, and the normal word stem process previously shown in Figs. 6 and 8. For all three subject areas it is seen that the word stem associations improve the recall values for the last few documents retrieved, over and above the values obtainable with the simple word stem matching process.

As an example of the performance of the concept-concept associations, consider search request QB2, titled "testing automated information systems", used with the ADI collection. One of the documents in this collection, number 80X, dealing with "experiments on documentation techniques" is relevant to the request, but is ranked only 77^{th} out of 82 for the regular word stem process, because very few of the words used in the document match the terms of the request. If concept-concept associations are generated, additional related terms such as "efficient", "real", "reduce", "experimental", "frequency", etc. are generated; these added terms provide a bridge between "test" and "experiments", and between "information" and "documentation", thus accounting for the improved performance.

While word-word correlations improve the basic word-stem matching process for high recall values, Fig. 12 shows that a well-constructed thesaurus is more powerful than the associative techniques applied to words. In other words, the thesaurus which serves much the same purpose as the associative process does so more accurately. This leads to the following conclusion:

Rule 7 : Statistical concept-concept associations can be used to improve recall performance particularly for collections for which a well ordered synonym dictionary does not exist.



Comparison of Thesaurus Performance with Statistical Word-Word Associations (Null Concon)

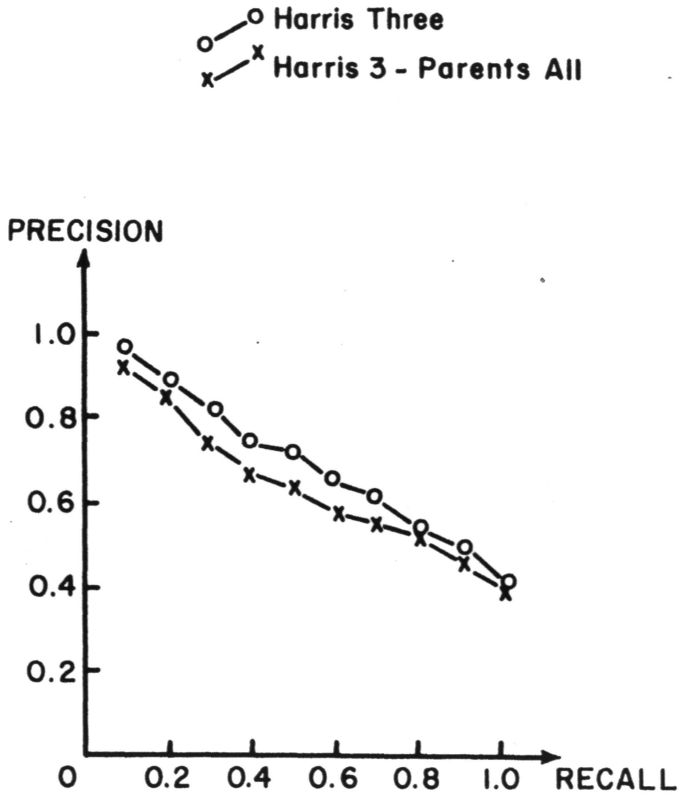
Fig. 12

E) Hierarchical Subject Expansion

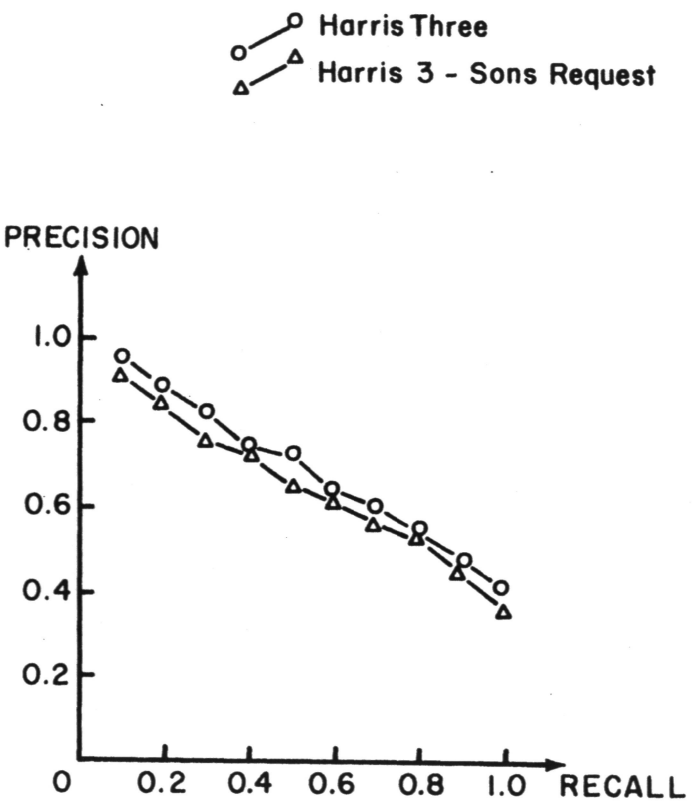
Hierarchical arrangements of information identifiers, similar in construction to library classification schedules make it possible, given an entry, to find more general terms by going "up" in the hierarchy (expansion by parents), and more specific ones by going "down" (expansion by sons). The hierarchies provided for the SMART system include, in addition, expansions by "brothers" on the same level as the original terms, and expansions by adding certain "cross-references". Dozens of different hierarchy options can be used, of which two are shown in Fig. 13.

Fig. 13(a) shows an expansion by adding for each original term its parent in the hierarchy, the expansion being applied to both documents and requests. Clearly, this option does not on the average provide an improvement over the standard "Harris Three" thesaurus process. On the other hand, an expansion by "sons" applied to requests only (and not to the documents) seems to offer some improvement in performance for the middle ranges of recall and precision.

In general, hierarchical subject expansions result in large-scale disturbances in the information identifiers attached to documents and search requests. Occasionally, such a disturbance can serve to crystallize the meaning of a poorly stated request, particularly if the request is far removed from the principal subjects covered by the document collection. More often, the change in direction specified by the hierarchy option is too violent, and the average performance of most hierarchy procedures does not appear to be sufficiently promising to advocate their incorporation in an analysis system for automatic document retrieval.



a) Hierarchy Expansion by Parents



b) Hierarchy Expansion by Sons

Sample Hierarchy Procedures (IRE-2 17 Requests)

Fig. 13

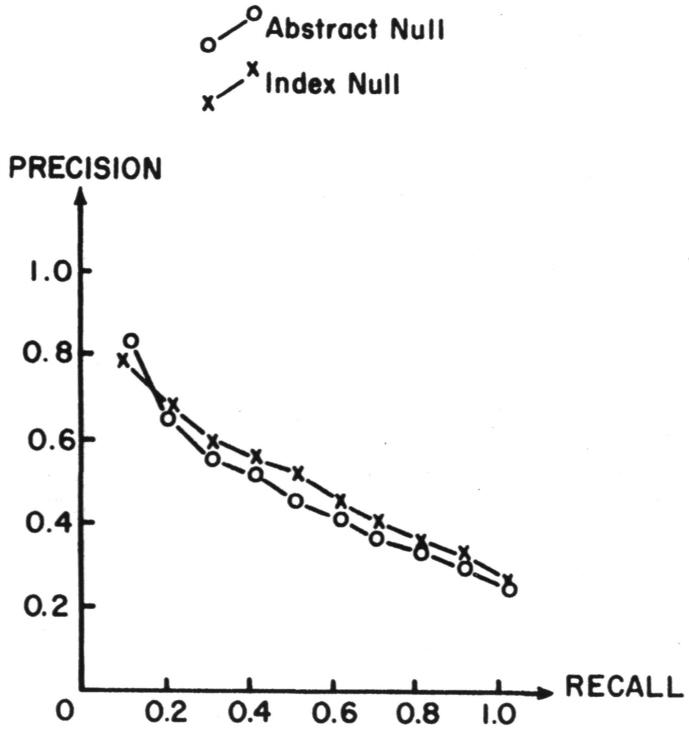
F) Manual Indexing

The Cranfield collections were available for purposes of experimentation both in the form of abstracts and in the form of manually assigned index terms. The indexing performed by subject experts is extremely detailed, consisting for some documents of over fifty index terms. As such, the indexing performance may be expected to be superior to the subject indexing normally used for large document collections. Nevertheless the output of Fig. 14(a) shows that the retrieval results obtained by matching the index terms ("index null") is only slightly superior to the standard word stem matching procedure, using the words extracted from the document abstracts.

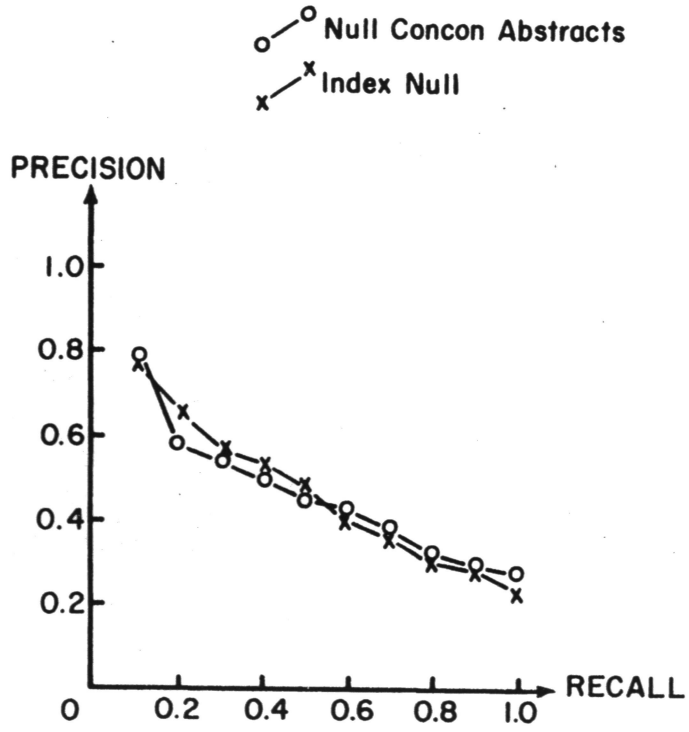
When the manual indexing procedure is compared with the word stem association process, it is seen in Fig. 14(b) that the word stem match with the associated terms is superior to the index term method. The same is true when manual indexing is compared with the regular thesaurus process. The output produced with the Cranfield collection then leads to the following rule:

Rule 8 : Keyword matching systems based on manually assigned index terms are found (at least for one well-known document collection) to be not substantially superior to raw word matching techniques, and to be actually inferior to statistical word association and to thesaurus methods.

This rule is in complete contradiction to what one hears repeated over and over again by documentation and library science specialists. Moreover, as the collection sizes increase, the manual indexing procedure



a) Manual Indexing vs Word Stem Process



b) Manual Indexing vs Word-Word Associations

Comparison of Manual Indexing with Text Processing

(Cranfield Collection)

Fig. 14

may be expected to decrease in effectiveness, because of the variabilities among indexers, and the difficulties of ensuring a uniform application of a given set of indexing rules to all documents. The computer process will, however, not decay as the collections grow larger, and one may anticipate for large collections of operational size an even greater difference in performance, and a clearer advantage for the automatic process.

G) Iterative Searching

Most presently operating information systems perform a single search operation for each search request, and the user of the system must submit a completely new request if he is dissatisfied with the initial response. This situation is not ideal, since it assumes that a single information analysis and search method will prove equally useful to all customers, and furthermore that all users have the same type of need and will thus be satisfied with the same type of answer. In actual practice, users have many different needs, some wanting very exhaustive answers, others being content with a single reference.

This situation is well recognized, and it is widely felt that the new computer time-sharing organizations, which permit a multiplicity of users to obtain access, more or less simultaneously, to a central equipment complex can be used advantageously to provide individualized service to each customer according to his need. Accordingly, several iterative search methods have been simulated with the SMART programs.[6,7] In each case, a user first obtains some output in response to an initial query, and depending on what he learns from this output, he returns enough information to the system to permit a reprocessing of the original query

under altered conditions.

The most effective procedure tried so far is the "relevance feedback" process, in which the user returns to the system a list of document numbers previously retrieved, together with information concerning the usefulness of each document for his search purpose. The system then automatically adjusts the original query by increasing the weight of query terms originally contained in documents identified as relevant, and simultaneously decreasing the weight of query terms contained in the nonrelevant document set. This process can, of course, be repeated several times, and results each time in a modification of the query in the "direction" of the document set termed relevant, and away from the document set termed nonrelevant.

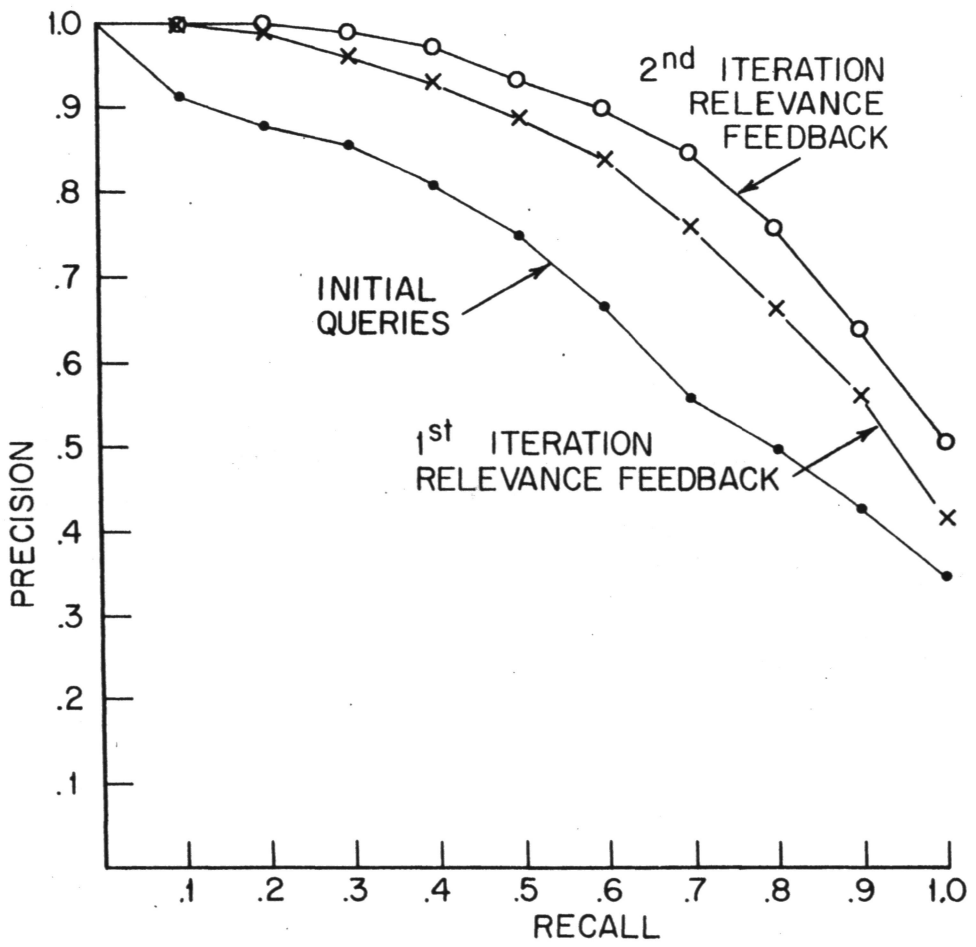
The results of two iterations performed with 24 search requests processed against the IRE - 2 collection are shown in Fig. 15. The first step of query modification is seen to result in a large-scale improvement in retrieval effectiveness, while the second iteration provides a smaller, but still pronounced increase in effectiveness.

Realistic tests of iterative search techniques can only be made in a real-time environment with adequate time-sharing equipment. The initial tests performed so far do, however, suggest the following rule:

Rule 9 : Iterative search techniques, based on feedback information supplied by the user as a result of previous retrieval procedures, appear to offer major promise for more effective search operations.

H) Summary

The principal conclusions resulting from the tests conducted with the SMART system are summarized in Fig. 16. These results suggest that



COMPARISON OF SIMPLE SEARCH WITH ITERATED SEARCH PROCESS
(IRE – 2 AVERAGES OVER 24 REQUESTS)

Fig. 15

1. Term Weights		
Weighted Word Stems	>>	Logical Stems
Weighted Synonym Classes	>>	Logical Synonym Classes
2. Document Length		
Full Summaries (2000 words)	>	Abstracts (150 words)
Abstracts (150 words)	>>	Titles Only
3. Synonym Recognition:		
Abstracts with Thesaurus	>>	Abstracts Null
Summaries with Thesaurus	>	Summaries Null
4. Phrase Recognition:		
Synonym and Phrase Recognition	>	Synonym Recognition (Thesaurus) only
5. Syntactic Analysis:		
Syntactic Analysis	>>	Word Stem Match
Syntactic Analysis	>	Synonym Recognition
Syntactic Analysis	~	Statistical Phrase Recognition
6. Term-Term Associations:		
Stem-Stem Associations	>	Simple Word Stems
Concept-Concept (Thesaurus Class) Associations	~	Synonym Recognition
7. Manual Indexing:		
Abstract Stem Matching	~	Index Term Match
Index Term with Thesaurus	>	Abstracts with Thesarus

>> "much greater than"
> "greater than"
~ "about equal to"

Overall Evaluation Results

(based on experiments with 4 collections in 3 topic areas)

Fig. 16

future information centers will make use of automatic text analysis rather than manual subject indexing. Among the techniques likely to be implemented in practice are the synonym recognition and phrase generation methods made possible by thesauruses and phrase dictionaries, and the statistical term-term association procedures. Document identifiers may be expected to be based on document abstracts, or longer document excerpts, and weights will be assigned to improve retrieval performance. A variety of additional techniques including expansion by subject hierarchies and automatic syntactic analyses may be used under special circumstances but their general applicability is still unproved.

Acknowledgement: The assistance of Mr. Cyril Cleverdon and Mr. Michael Keen of the Aslib-Cranfield Research Project in making available the Cranfield documents and dictionaries is gratefully acknowledged.

References

[1] Committee on Scientific and Technical Information (COSATI), Recommendations for National Document Handling System, Report PB 168267 distributed by National Clearinghouse, November 1965.

[2] M. Rubinoff, editor, Toward a National Information System, Spartan Books, Washington 1965.

[3] G. Salton, A Document Retrieval System for Man-machine Interaction, Proceedings of the ACM 19th National Conference, Philadelphia, Pa., 1964.

[4] G. Salton and M. E. Lesk, The SMART Automatic Document Retrieval System -- An Illustration, Communications of the ACM, Vol. 8, No. 6, June 1965.

[5] G. Salton, Progress in Automatic Information Retrieval, IEEE Spectrum, Vol. 2, No. 8, August 1965.

[6] J. J. Rocchio and G. Salton, Information Search Optimization and Iterative Retrieval Techniques, Proceedings of the Fall Joint Computer Conference, Las Vegas, November 1965.

[7] J. J. Rocchio, Document Retrieval Systems -- Optimization and Evaluation, Doctoral Thesis, Report No. ISR-10 to the National Science Foundation, Computation Laboratory, Harvard University, April 1966.

[8] C. W. Cleverdon, The Testing of Index Language Devices, Aslib Proceedings, Vol. 15, No. 4, April 1963.

[9] G. Salton, The Evaluation of Automatic Retrieval Procedures -- Selected Test Results Using the SMART System, American Documentation, Vol. 16, No. 3, July 1965.

[10] G. Salton, The Evaluation of Automatic Information Systems, 1965 International FID Congress, Washington, October 1965.

[11] R. A. Fairthorne, Basic Parameters of Retrieval Tests, 1964 ADI Annual Meeting, Philadelphia, October 1964.

[12] C. Cleverdon, J. Mills and M. Keen, Factors Determining the Performance of Indexing Systems, Vol. 1-Design, Cranfield, 1966.

[13] G. Salton, Information Dissemination and Automatic Information Systems, to be published Proceedings of the IEEE, December 1966.

References (contd.)

[14] G. Salton, Automatic Phrase Matching, in Readings in Automatic Language Processing, D. Hays, editor, American Elsevier, New York 1966.

[15] L. B. Doyle, Indexing and Abstracting by Association, American Documentation, Vol. 13, No. 4, October 1962.

[16] V. E. Giuliano and P. E. Jones, Linear Associative Information Retrieval, in Vistas in Information Handling, P. Howerton, editor, Spartan Books, Washington, D. C., 1963.