III. Computer Evaluation of Indexing and Text Processing

G. Salton and M. E. Lesk

Automatic indexing methods are evaluated and design criteria for modern information systems are derived.

1. Introduction

Throughout the technical world, a growing interest is evident in the design and implementation of mechanized information systems. Over the last few years, the general feeling that something should be done to help organize and store some of the available information resources has given way to the widespread impression that modern computing equipment may in fact be capable of alleviating and solving to some extent the so-called information problem. Specifically, it is believed that the required capacity exists to store many data or document collections of interest, that procedures are available for analyzing and organizing the information in storage, and that realtime software and hardware can be used to insure that the stored information is retrieved in response to requests from a given user population in a convenient form, and at little cost in time and effort. [1,2,3]

Before investing the necessary resources required for the implementation of sophisticated information services, it becomes necessary to generate the detailed systems specifications and to determine which of many possible alternative design features should in fact be implemented. This, in turn, must be made to depend on experimentation in a controlled environment to test and evaluate the effectiveness of various possible search and analysis procedures. The SMART document retrieval system which has been operating on an IBM 7094 for over 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 iteration with the system. [4,5,6,7]

The present study summarizes the results obtained with the SMART system over a two year period starting in 1964, and presents evaluation output based on the processing of three document collections in three different subject fields. Conclusions are drawn concerning the most likely analysis methods to be implemented in an operational environment. The emphasis throughout is on text analysis procedures since they form an important part of a document handling system. Several operational problems, including the actual network implementation of a retrieval system are not covered; cost and timing estimates are also excluded, because these are tied directly to the specific environment within which a given system actually operates.

The basic features of the SMART system are first described, and the design of the main experiments is outlined, including the statistical procedures used to test the significance of the evaluation output obtained. The principal evaluation results are then presented, and tentative conclusions are reached concerning the effectiveness of automatic text analysis procedures as part of future information systems. The results derived from the present experiments are also briefly compared with the output obtained with several other testing systems.

2. The SMART System

A) Basic Organization

The SMART system 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, phrase generating methods, and the like, in order to obtain the content identifications useful for the retrieval process.

The following facilities incorporated into the SMART system for purposes of document analysis are of principal interest:

a) a system for separating English words into stems and affixes which can be used to reduce incoming texts into word stem form;

b) a synonym dictionary, or thesaurus, used to replace significant word stems by concept numbers, each concept representing a class of related word stems;

c) a hierarchical arrangement of the concepts included in the thesaurus which makes it possible, given any concept number, to find its "parent" in the hierarchy, its "sons", its "brothers", and any of a set of possible cross-references;

d) statistical association methods used to compute similarity coefficients between words, word stems, or concepts, based on cooccurrence patterns between these entities in the sentences of a document, or in the documents of a collection; associated

items can then serve as content identifiers in addition to the original ones;

e) syntactic analysis methods which permit the recognition and use as indicators of document content of phrases consisting of several words or concepts where each element of a phrase must hold a specified syntactic relation to each other element;

f) statistical phrase recognition methods which operate like the preceding syntactic procedures by using a preconstructed phrase dictionary, except that no test is made to ensure that the syntactic relationships between phrase components are satisfied;

g) request-document matching procedures which make it possible to use a variety of different correlation methods to compare analyzed documents with analyzed requests, including concept weight adjustments and variations in the length of the document texts being analyzed.

Stored documents and search requests are processed by the system without any prior manual analysis using 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. [4,5,6] A cut-off can then be picked, and documents above the chosen cut-off can be withdrawn from the file and turned over to the user as answers to the search request.

The search process may be controlled by the user in that a request can be processed first in a standard mode. After analysis of the output produced, feedback information can then be returned to the system where it is used to reprocess the request under altered conditions. The

new output can again be examined, and the search can be iterated until the right kind and amount of information are obtained.[7,8]

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 evaluation procedures actually used are described in the next section.

B) Evaluation Process

The evaluation of an information search and retrieval system can be carried out in many different ways depending on the type of system considered — whether operational, experimental with user populations, or laboratory type system, on the viewpoint taken — that of the user, the manager, or the operator, and on other factors, such as the special aims of the evaluation study. A large number of different variables may affect the results of any evaluation process, including the kind of user population, the type and coverage of the document collection, and indexing tools used, the analysis and search methods incorporated into the system, the input-output equipment used, the operating efficiency as well as costs and time lag needed to produce answers, and many others.

In the present context, the viewpoint taken is the user's and the overriding criterion of systems effectiveness is taken to be the ability of the

system to satisfy the user's information need. Management criteria such as cost are not taken into account, even though in the final analysis the problem is of primary importance, since the most effective system will not avail if the operations are too costly to be performed. However, costs are difficult to measure in an experimental situation where unusual fluctuations may occur because of many extraneous factors. Furthermore, the immediate need is for a measurement of the effectiveness of the intellectual tools used to analyze and search the stored information, since these are responsible in large part for the retrieval results. Costs can later be taken into account, for example, by providing several classes of service at varying cost.

The evaluation measures actually used are based on the standard recall 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. In an operational situation, where information needs may vary from user to user, some customers may require high recall — that is the retrieval of most everything that is likely to be of interest — while others may prefer high precision — that is, the rejection of everything likely to be useless. Everything else being equal, a perfect system is one which exhibits both a high recall and a high precision.

If a cut is made through the document collection to distinguish retrieved items from nonretrieved on the one hand, and if procedures are available for separating relevant items from nonrelevant ones on the other, the standard recall R and standard precision P may be defined as follows:

$$R = \frac{\text{number of items retrieved and relevant}}{\text{total relevant in collection}} ,$$

and

$$P = \frac{\text{number of items retrieved and relevant}}{\text{total retrieved in collection}} .$$

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]

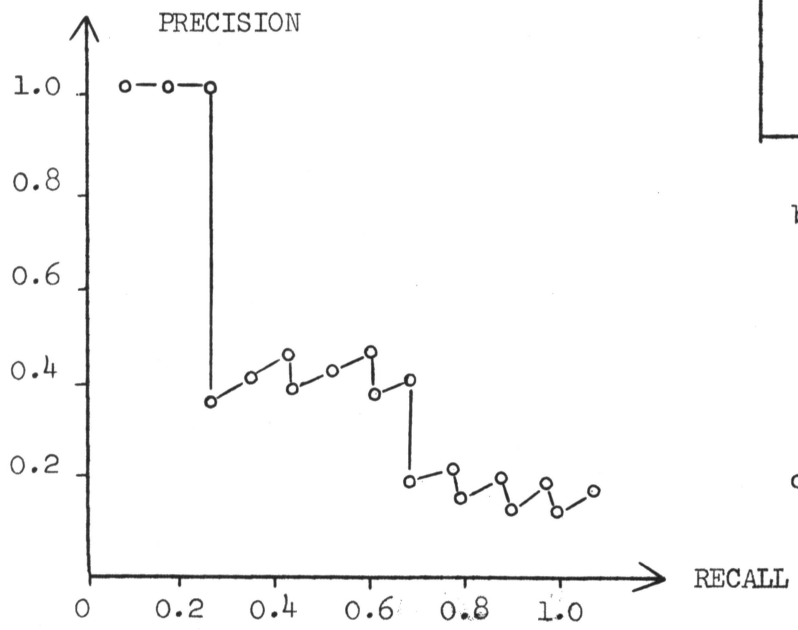
In the evaluation work carried out with the SMART system, manually derived, exhaustive relevance judgments are used since the document collections processed are all relatively small. Moreover, the choice of a unique cut-off is avoided by computing the precision for various recall values, and exhibiting a plot showing recall against precision. An example, of such a graph is shown in Fig. 1 for query Q145, processed against a collection of 200 documents in aerodynamics. A total of twelve documents in the collection were judged relevant to the request, the relevance judgments being performed by a subject expert independently of the retrieval system. The ranks of the relevant documents produced by the search system after ordering of the documents in decreasing correlation order are shown in Fig. 1(a). For the retrieval process illustrated in Fig. 1, these ranks range from 1 for the relevant document with the highest request-document correlation to 78 for the relevant item with the lowest correlation. By choosing successive cut-off values after the retrieval of 1, 2, 3, ..., n documents, and computing recall and precision values at each point, a recall-precision table can be constructed, as shown in Fig. 1(b). The recall-precision graph obtained from this table is represented in Fig. 1(c).

Relevant Documents		
Rank	Number	Correlation
1	80	.5084
2	102	.4418
3	81	.4212
10	82	.2843
11	193	.2731
14	83	.2631
15	87	.2594
20	88	.2315
40	86	.1856
50	109	.1631
69	84	.1305
78	85	.1193

a) List of Relevant Documents

Recall-Precision after Retrieval of X Documents		
X	Recall	Precision
1	0.0833	1.0000
2	0.1667	1.0000
3	0.2500	1.0000
9	0.2500	0.3333
10	0.3333	0.4000
11	0.4167	0.4545
13	0.4167	0.3846
14	0.5000	0.4286
15	0.5833	0.4667
19	0.5833	0.3684
20	0.6667	0.4000
39	0.6667	0.2051
40	0.7500	0.2250
49	0.7500	0.1837
50	0.8333	0.2000
68	0.8333	0.1470
69	0.9167	0.1594
77	0.9167	0.1428
78	1.0000	0.1538

b) Recall Precision Table



c) Recall-Precision Plot

Performance Characteristics for Query Q145
(Cranfield-1, Word Stem Run)

Fig. 1

Recall-precision graphs, such as that of Fig. 1(c) have been criticized because a number of parameters are obscured when plotting recall against precision — for example, the size of the size of the retrieved document set and the collection size.[11] Such plots are, however, 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]

In addition to the standard recall and standard precision measures, whose value depends on the size of the retrieved document set, it is also possible to use indicators which are independent of the retrieved set. In particular, since the SMART system produces ranked document output in decreasing order of correlation between documents and search requests, evaluation measures can be generated which are based on the ranks of the set of relevant documents, as determined by the automatic retrieval process, compared with the ranks of the relevant documents for an ideal system where all relevant items are retrieved before any nonrelevant ones.

Two particularly attractive measures with this property are the normalized recall and normalized precision, which are defined as follows [7,9]:

$$R_{norm} = 1 - \frac{\sum_{i=1}^{n} r_i - \sum_{i=1}^{n} i}{n(N-n)} \quad ,$$

$$P_{norm} = 1 - \frac{\sum_{i=1}^{n} \log r_i - \sum_{i=1}^{n} \log i}{\log \frac{N!}{(N-n)\,!\,n!}}$$

where n is the size of the relevant document set, N is the size of the total document collection, and r_i is the rank of the i^{th} relevant document when the documents are arranged in decreasing order of their correlation with the search request.

These measures range from 1 for a perfect system in which all relevant items are placed at the top of the retrieved list, to 0 for the worst case where all nonrelevant items are retrieved before any relevant one. Furthermore, under certain circumstances the normalized measures can be shown to be closely related to the standard measures as follows[12]:

$$R_{norm} \approx \frac{1}{N} \sum_{i=1}^{n} R(i)$$

when the number of relevant documents n is small compared to the collection size N, and

$$P_{norm} \approx \frac{1}{N} \sum_{i=1}^{n} P(i)$$

for large N and n not too small. $R(i)$ and $P(i)$ correspond, respectively, to the standard recall and precision values after the retrieval of i documents.

Two further overall measures of retrieval effectiveness, analogous to the normalized measures, but somewhat simpler to compute, are the "rank recall" and "log precision" measures, defined as

$$\text{rank recall} = \frac{\sum_{i=1}^{n} i}{\sum_{i=1}^{n} r_i},$$

and

$$\text{log precision} = \frac{\sum_{i=1}^{n} \ln i}{\sum_{i=1}^{n} \ln r_i}$$

where n is again equal to the number of relevant documents, and r_i is the rank (in decreasing correlation order) of the i^{th} relevant document. Like the normalized measures, rank recall and log precision are functions of the rank of the relevant documents, but contrary to the earlier situation, these measures do not take into account the collection size N.

Under normal circumstances, the results of a systems evaluation must reflect overall system performance, rather than the performance for individual requests only. In these circumstances, it is convenient to process many search requests and to use an average performance value as a measure of retrieval effectiveness. [12] For the overall evaluation measures (normalized recall) and precision, and rank recall and log precision), the averaging process presents no

problem, since only a single set of values is obtained in each case for each request. The averaging method is more complex for the standard recall-precision graph, since a continuous set of values is involved, and the number of relevant documents differs from request to request. In the SMART system, the averaging process for the recall-precision graph corresponding to many search requests is performed as follows:

a) ten specified standard recall values are picked ranging from 0.1 to 1.0;

b) for each recall level, the number of documents which must be retrieved in order to obtain the specified level is determined;

c) using the cut-off value thus calculated for the number of retrieved documents, the precision value is generated corresponding to the specified recall;

d) the precision values obtained for a given recall value are averaged over a number of search requests, and the resulting point is added to the recall-precision plot;

e) the ten individual points on the plot are joined to produce an average recall-precision curve.

Averaged evaluation results are presented for three different document collections in section 3.

C) Significance Computations

For each search request and each processing method, the evaluation procedure incorporated into the SMART system produces fourteen different statistics, including four global statistics (rank recall, log

precision, normalized recall and normalized precision), and ten local statistics (standard precision for ten recall levels). A problem then arises concerning the use of these fourteen statistics for the assessment of systems performance. In theory, comparisons between different processing methods are easy to make by contrasting, for example, the recall-precision plots obtained in each case. In practice, it is difficult to draw hard conclusions because the variation in performance between individual requests is large, because the fourteen measures have different ranges, and because it is unclear a priori whether the magnitude of a given recall or precision value is significant or not. In particular, given a specified recall or precision value, it is of interest to determine whether values as large, or larger, as the given one could be expected under random circumstances, or whether on the contrary the probability of obtaining a given specified value for an average system is very small.

Because of the large request variance and the differences in the range of the various parameters, the significance computations incorporated into the SMART system are based on paired comparisons between the request performance using processing method A, and the performance using method B. In particular, the difference in magnitude is computed for each of the fourteen pairs of statistics obtained for each request for each pair of processing methods. These differences are then averaged over many requests, and statistical computations are used to transform the averaged differences into probability measurements. Each of the fourteen values thus obtained represents the probability that if the performance level of the two methods A and B were in fact equally high for the given statistic (except for random variations in the test results), then a test value as large as the one actually observed would occur for a system. A probability

value of 0.05 is usually taken as an upper bound in judging whether a deviation in test values is significant or not. Using this probability value as a limit, the corresponding test difference would in fact be significant nineteen times out of twenty, and only one time out of twenty would two equally effective systems be expected to produce as large a test difference.

Since it is difficult to judge systems performance by using fourteen different probability values, corresponding to the fourteen evaluation measures, an aggregate probability value is computed from the fourteen individual probabilities. The significance of this aggregate depends on the independence of the various significant tests.

Two separate testing procedures are incorporated into the SMART system. The first one uses the well-known t-test based on Student's t-distribution. [13] This test requires an underlying normal distribution of the data used in the test process, as well as the independence among the search requests processed against the document collections. The t-test process takes into account the actual magnitude of the differences for the statistics being calculated, and the resulting probabilities are considered to be reliable indicators of system differences.

A less demanding testing procedure is furnished by the sign-test, where the magnitude of the differences in the statistics is not taken into account, but only the sign of the differences (that is, an indication of whether method A provides a larger test result than B, or

vice-versa). [14] An attractive feature of the sign-test is that normality of the input data is not required, and since this normality is generally hard to prove for statistics derived from a request-document correlation process, the sign test probabilities may provide a better indicator of system performance than the t-test.

The t-test computations are performed as follows: let m_{ijA} be the value of statistic i for request j, using method A (for example, the value of the rank recall, or the normalized recall). Then, given two processing methods A and B, and a set of k requests, the average of the differences for statistic i are computed. Specifically,

$$d_{ij} = m_{ijA} - m_{ijB} ,$$

and

$$D_i = \frac{1}{K} \sum_{j=1}^{k} d_{ij} .$$

The difference computations for two statistics (rank recall and log precision) are shown in Fig. 2. The average differences are then used to obtain the standard deviation of the differences $(SD)_i$ and the t-test values T_i, where $T_i = (D_i / (SD)_i) \cdot k$.

The t-test values T_i are now converted to probabilities P_{ti} using Student's t-distribution with k degrees of freedom.

The probabilities derived from the fourteen statistics are then used to compute an aggregate probability by first converting the two tailed t-test

Request Name	Rank Recall		Difference	Log Precision		Difference
	Method A	Method B		Method A	Method B	
AUTOMATA PHR	0.5238	0.9649	-0.4411	0.7126	0.9881	-0.2755
COMP SYSTEMS	0.0725	0.1228	-0.0503	0.3783	0.4806	-0.1023
COMPS-ASSEMB	0.3714	0.7428	-0.3714	0.8542	0.9453	-0.0911
CORE MEMORY	0.0691	0.1064	-0.0373	0.3157	0.3695	-0.0538
DIFFERNTL EQ	0.5298	0.7574	-0.2276	0.8620	0.9219	-0.0599
ERROR CONTRL	0.1460	0.1875	-0.0415	0.5342	0.5972	-0.0630
M10-COUNTERS	0.8182	0.7347	0.0835	0.8682	0.8599	0.0083
M2TRANSMIT	0.0522	0.0963	-0.0441	0.2819	0.4698	-0.1879
M3-INFORM	0.1968	0.3134	-0.1166	0.6300	0.7666	-0.1366
M8-STORAGE	0.0375	0.2763	-0.2388	0.2670	0.4666	-0.1996
MISSILE TRAK	1.0000	0.7500	0.2500	1.0000	0.6309	0.3691
MORSE CODE	1.0000	1.0000	0.0000	1.0000	1.0000	0.0000
PATTERN RECG	1.0000	1.0000	0.0000	1.0000	1.0000	0.0000
RANDOM NUMBS	0.0517	0.2000	-0.1483	0.1750	0.3408	-0.1658
SOLSTAT CIRC	0.2766	0.3402	-0.0636	0.6921	0.7912	-0.0991
SWITCH FUNCS	0.3529	0.4444	-0.0915	0.7416	0.8005	-0.0589
THIN FILMS	0.2157	0.8462	-0.6305	0.6294	0.9242	-0.2948
Total	6.7142	8.8833	-2.1691	10.9422	12.3531	-1.4109
Average Value over 17 Requests	0.3950	0.5225	-0.1276	0.6437	0.7267	-0.0830

Computation of Recall and Precision Differences
for Individual Requests

(Method A: Stem Concon; Method B: Thesaurus)

Fig. 2

	Evaluation Measure		Average Value: Method A	Average Value: Method B	Difference of Average	Standard Deviation	t-Test Value	Probability
1	Rank Recall		0.3950	0.5225	-0.1276	2.07E-01	2.54E 00	0.0219
2	Log Precision		0.6437	0.7267	-0.0830	1.47E-01	2.33E 00	0.0334
3	Normed Recall		0.9233	0.9675	-0.0442	5.35E-02	3.41E 00	0.0036
4	Normed Precision		0.7419	0.8639	-0.1219	1.20E-01	4.19E 00	0.0007
5	Precision Graph (Precision for ten Recall Levels)	0.1	0.7385	0.9735	-0.2351	2.88E-01	3.37E 00	0.0039
6		0.2	0.6544	0.8973	-0.2428	2.82E-01	3.55E 00	0.0026
7		0.3	0.5844	0.8245	-0.2401	2.51E-01	3.95E 00	0.0011
8		0.4	0.5326	0.7551	-0.2226	2.39E-01	3.84E 00	0.0014
9		0.5	0.5187	0.7146	-0.1959	2.00E-01	4.04E 00	0.0009
10		0.6	0.5035	0.6499	-0.1464	1.59E-01	3.79E 00	0.0016
11		0.7	0.4452	0.6012	-0.1561	1.79E-01	3.59E 00	0.0024
12		0.8	0.4091	0.5514	-0.1423	2.24E-01	2.62E 00	0.0184
13		0.9	0.3794	0.4973	-0.1179	2.29E-01	2.12E 00	0.0499
14		1.0	0.3106	0.4118	-0.1012	2.44E-01	1.71E 00	0.1070

Combined Significance: Total Chi-Square 1.67E 02
Total Probability of B over A 0.0000

t-Test Computations for 14 Different Recall and Precision Measures

(Averages over 17 requests, Method A: Stem Concon,
Method B: Thesaurus)

Fig. 3

to a one-tailed test, changing each probability to chi-square, adding the chi-square values, and finally reconverting to a probability P_t, using a chi-square distribution with 28 degrees of freedom.[13] Specifically, let s be the sign of the sum of the differences D_i, or

$$s = \text{sign}\left(\sum_i D_i\right).$$

Then if sign $D_i = s \Rightarrow P'_{ti} = \frac{1}{2}P_{ti}$; alternatively

if sign $D_i \neq s \Rightarrow P'_{ti} = 1 - \frac{1}{2}P_{ti}$.

The chi-square of the sum is now obtained such that

$$X^2 = -\sum_{i=1}^{14} - 2 \log P'_{ti} .$$

Finally, this value is converted to the desired probability P_t.

The t-test computations are shown for two sample analysis methods A and B in Fig. 3. The values in the first two columns of Fig. 3 represent averages over 17 search requests for each of the fourteen evaluation measures. The final probabilities P_{ti} range from a high of 0.107 for the standard precision at recall value 1, to a low of 0.0007 for the normalized precision. The final probability value P_t is smaller than 1.10^{-4}, thus indicating that the combination algorithm concentrates on the significant tests, while ignoring the less significant ones. The validity of the process depends on an assumption of independence among the fourteen measures, which is true to a limited extent for the measures used.

The sign-test uses the binomial instead of the t-distribution to produce a probability value. Specifically, given two processing methods for which the null hypothesis applies (that is, two equivalent methods), each d_{ij} has equal chances of being positive or negative; moreover, since the search requests are assumed unrelated (independent) and randomly distributed, the signs of the differences are unrelated. The number, say M, of positive signs is accordingly binomially distributed with p equal to one-half and k equal to the number of requests.

M can then serve as a statistic to test the null hypothesis by taking large values of M as significant evidence against the equivalence of the two methods tested. Obviously, a test based on rejecting the equivalence hypothesis for large values of M, is equivalent to one based on rejection for small values of M', the number of negative signs. As before, a probability of 0.05 may be taken as an upper limit for rejecting the equivalence assumption.

Since the sign test does not depend on the magnitudes of the differences, the number of positive, or negative signs can be cumulated directly. In particular, the number of requests preferring method A is summed over all measures, as well as the number of requests preferring method B. These totals are then subjected to the same testing process, as follows: let t be a tolerance value, taken as 0.001 for the present test; further, for each statistic i, let

k_{ai} be the number of d_{ij} $(j=1,....k)$ exceeding $+t$,

k_{bi}, the number of d_{ij} smaller than $-t$,

and k_{ci}, the number of d_{ij} such that $|d_{ij}| \leq t$,

where the number of requests $k = k_{ai} + k_{bi} + k_{ci}$.

The sign-test probability for statistic i is now computed as follows: let $k_{vi} = k_{ai} + k_{bi}$,

and $k_{wi} = \min (k_{ai}, k_{bi})$,

then $$P_{si} = \sum_{j=1}^{k_{wi}} \frac{k_{vi}!}{j!(k_{vi}-j)!} \; 2^{-k_{vi}+1} \quad .$$

The overall probability, P_s, can be cumulated directly for the fourteen evaluation measures; specifically,

$$\text{if } k_a = \sum_i k_{ai} \quad ,$$

$$k_b = \sum_i k_{bi} \quad ,$$

$$k_v = k_a + k_b \quad ,$$

$$\text{and } k_w = \min (k_a, k_b),$$

$$\text{then } P_s = \sum_{j=1}^{k_w} \frac{(k_v)!}{j!(k_v-j)!} \; 2^{-k_v+1} \quad .$$

The sign test computations are shown in Fig. 4 for the same processing methods and search requests previously used as examples in Figs. 2 and 3. The individual probabilities P_{si} range in values from 0.0010 to 0.1185. The overall probability is again smaller than 1.10^{-4}; this is also reflected by the fact that method B is preferred 165 times, while A is superior only 26 times, with 47 ties.

Evaluation Measure	Number of Requests Superior for Method A	Number of Requests Superior for Method B	Number of Request Equal for A and B	Probability (B over A)
Rank Recall	2	13	2	0.0074
Log Precision	2	13	2	0.0074
Normed Recall	2	13	2	0.0074
Normed Precision	2	13	2	0.0074
Recall-Precision Graph 0.1	0	9	8	0.0039
0.2	0	11	6	0.0010
0.3	1	12	4	0.0034
0.4	1	11	5	0.0036
0.5	0	13	4	0.0002
0.6	3	11	3	0.0574
0.7	2	12	3	0.0129
0.8	3	12	2	0.0352
0.9	4	11	2	0.1185
1.0	4	11	2	0.1185
Combined Significance for 14 Measures	26	165	47	0.0000

Sign Test Computations for 14 Different Recall and Precision Measures

(Averages over 17 Requests; Method A: Stem Concon,
Method B: Thesaurus)

Fig. 4

Since the fourteen statistics used may not be fully independent, a question arises concerning the interpretation of the cumulated t-test probability P_t, and the cumulated sign test probability P_s. As a general rule, the equality hypothesis between two given methods A and B can safely be rejected when both probabilities P_s and P_t do not exceed 0.001 in magnitude, implying that most of the individual probabilities P_{si} and P_{ti} are smaller than 0.05, and when the same test results are obtained for all document collections being tested. If, on the other hand, the values of the final probabilities are larger, or if the test results differ from one collection to the next, additional tests would seem to be required before a decision can be made.

3. Experimental Results

A) Test Environment

The principal parameters controlling the test procedure are listed in Figs. 5, 6, and 7, respectively. The main properties of the document collections and search requests are shown in Fig. 5. Specifically, results are given for three document collections in the following subject fields:

a) Computer Science (IRE-3): a set of 780 abstracts of documents in the computer literature, published in 1959-1961, and used with 34 search requests;

b) Documentation (ADI): a set of 82 short papers, each an average of 1380 words in length, presented at the 1963 Annual Meeting of the American Documentation Institute, and used with 35 search requests;

c) Aerodynamics (CRAN-1): a set of 200 abstracts of documents used by the second Aslib Cranfield Project [15], and used with 42 search requests.

Characteristics			IRE-3	CRAN-1	ADI
Document Collection	Number of documents in collection		780	200	82
	Average number of words (all words) per document	full text	-	-	1380
		abstract	88	165	59
		title	9	14	10
	Average number of words (common words deleted) per document	full text	-	-	710
		abstract	49	91	35
		title	5	11	7
	Average number of concepts per analyzed document	full text	-	-	369
		abstract	40	65	25
		title	5	9	6
Search Requests	Number of search requests		34	42	35
	Average number of words per request (all words)		22	17	14
	Request preparation				
	a) short paragraphs prepared by staff members for test purposes		✓		✓
	b) short paragraphs prepared by subject experts previously submitted to operational system			✓	

Document Collection and Request Characteristics

Fig. 5

Characteristics	IRE-3	CRAN-1	ADI
Preparation of relevance judgments			
a) dichotomous prepared by staff experts based on abstracts using full relevance assessment	✓		
b) dichotomous prepared by subject experts based on abstracts and full text (full relevance assessment)		✓	
c) dichotomous prepared by staff experts based on full text using full relevance assessment			✓
Number of relevant documents per request (all requests)			
a) range	2-65	1-12	1-33
b) mean	17.4	4.7	4.9
c) generality (mean divided by collection size)	22.2	23.6	59.2
Number of relevant documents per specific request			
a) number of specific requests	17	21	17
b) mean number of relevant	7.5	3.0	2.1
Number of relevant documents per general request			
a) number of general requests	18	21	18
b) mean number of relevant	25.8	6.4	7.4

Relevance Distribution and Assessment

Fig. 6

Characteristics	IRE-3	CRAN-1	ADI
User population			
a) 10 students and staff experts	✓		✓
b) 42 subject experts		✓	
Number of retrieved documents per request	all	all	all
Number of indexing and search programs used			
a) matching algorithms	2	2	2
b) term weight adjustment	2	2	2
c) document length variation	3	4	3
d) basic dictionaries (suffix 's', stem, thesaurus, stat. phrases, hierarchy, syntax)	6	5	5
e) concept-concept association dictionaries	2	3	1
f) total basic options	144	240	60

General Test Environment

Fig. 7

Each of these collections belongs to a distinct subject area, thus permitting the comparison of the various analysis and search procedures in several contexts. The ADI collection in documentation is of particular interest because full papers are available rather than only document abstracts. The Cranfield collection, on the other hand, is the only one which is also manually indexed by trained indexers, thus making it possible to perform a comparison of the standard keyword search procedures with the automatic text processing methods.

The procedure used to collect relevance assessments and the related statistical information concerning the average number of relevant documents per request are summarized in Fig. 6. Exhaustive procedures were used to assess the relevance of each document with respect to each search request. Only one person (the requestor) was asked to collect the judgments for each request, and dichotomous assessments were made to declare each document as either relevant or not. In the words of a recent study on evaluation methodology, the process used consists of "multiple events of private relevance".[17]

Additional data concerning the user population and the number of search programs used are given in Fig. 7. In each case, the user population consisted of volunteers who were asked to help in the test process. Several hundred analysis and search methods incorporated into the SMART system were used with the three document collections. Results based on about sixty of these processing methods are exhibited in the present study.

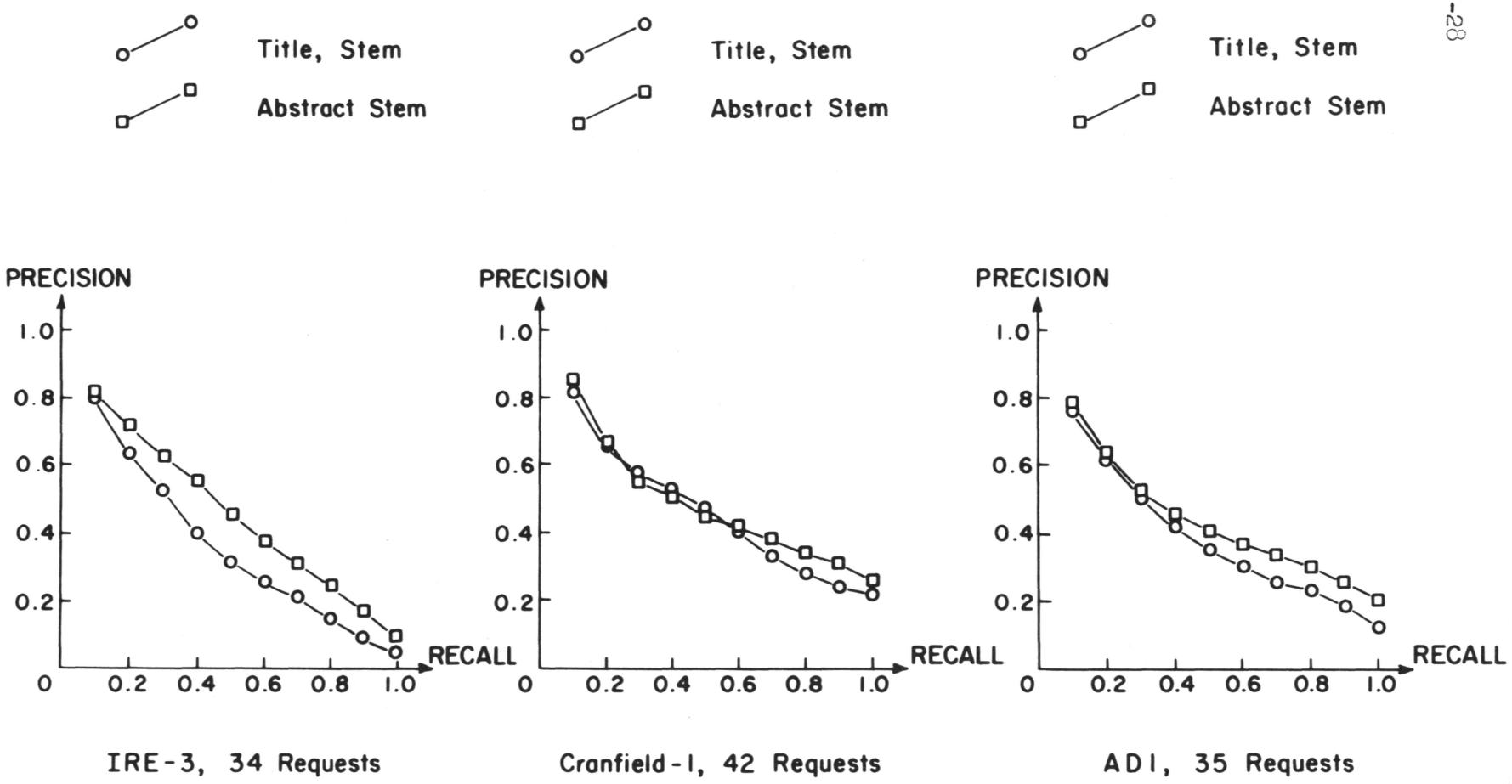
The methods chosen are generally useful to answer a number of basic

questions affecting the design of automatic information systems: for example, 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; what should be the role of the user in formulating and controlling the search procedure. These and other questions are considered in the evaluation process described in the remainder of this section.

B) Document Length

A first variable of interest is the length of each document to be used for content analysis purposes. This fundamental question enters into many of the arguments between advocates of automatic systems, and others who hold that manual content analysis methods are essential, because in an automatic environment, it is not normally possible to process the full text of all documents.

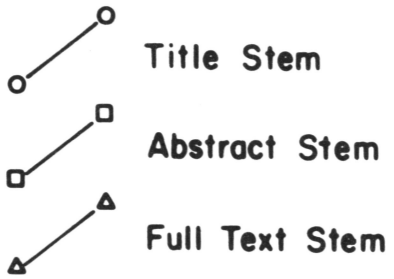
In Fig. 8, three analysis systems based on document titles only are compared with systems based on the manipulation of complete document abstracts. In each case, weighted word stem, extracted either from the titles or from the abstracts of the documents, are matched with equivalent indicators from the search requests. Fig. 8 exhibits recall-precision

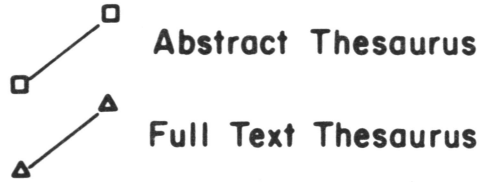


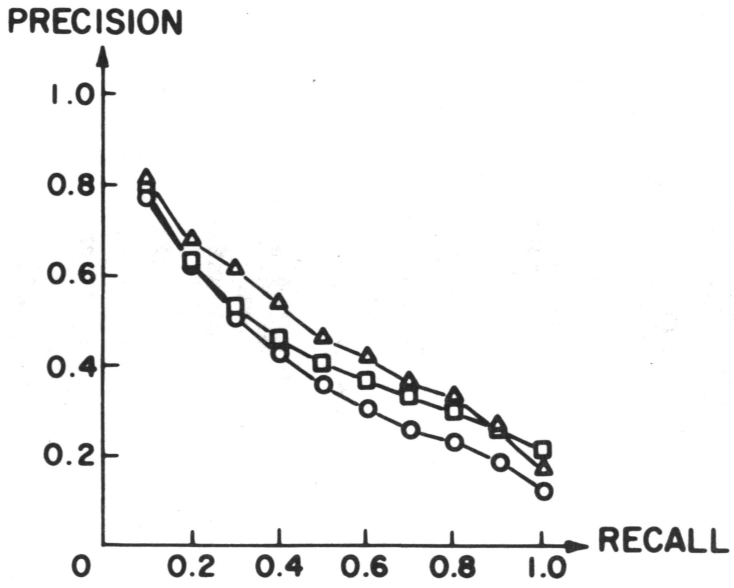
Comparison of document length

(Cosine correlation; numeric vectors)

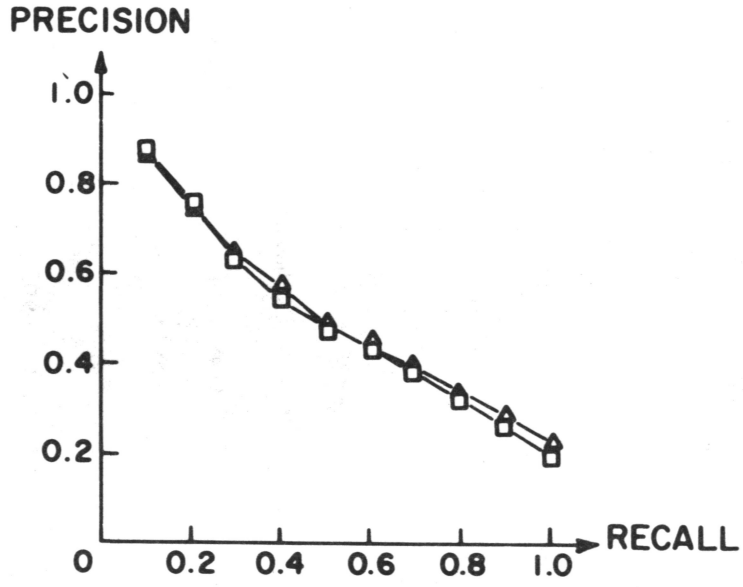
Fig. 8







ADI, 35 Requests



ADI, 35 Requests

Document Length (Full Text) Comparison

(Cosine correlation: numeric vectors)

graphs, averaged, respectively, over 34, 42, and 35 search requests for the computer science, aerodynamics, and documentation collections. In every case, the abstract process is found to be superior to the "title only" option, particularly at the high recall end of the curve, since the abstract curve comes closest to the upper righthand corner of the graph where both recall and precision are equal to 1. (For an ideal system which retrieves all relevant items before any irrelevant ones the recall-precision curve shrinks to a single point with coordinates (1,1)).

The significance output for the graphs shown in Fig. 8 to 22 is collected in Fig. 23. In each case, reference is made to the graphs being compared, and the combined probability values P_s and P_t are listed with an indicator specifying the preferred method. The superiority of the "abstract - stem" process of Fig. 8 is reflected in the significance output of Fig. 23. The probability of a correct null hypothesis is smaller than 1.10^{-4} for both the sign and the t-tests, thus showing that document titles are definitely inferior to document abstracts as a source of content indicators.

The ADI documentation collection was used to extend the analysis to longer document segments. Fig. 9 shows the results of a comparison between document abstract processing (60 words) and full text processing (1400 words), using both the word stem analysis, where weighted word stems are directly extracted from the texts and used as content indicators, and the thesaurus process, where synonymous word stems are first recognized by a dictionary look-up process before the document identifiers are

matched with the request identifiers. In each case, the full text process is superior to the abstract process, but the improvement in performance appears smaller than that shown in Fig. 8 for the title-abstract comparison.

The significance output for the tests of Fig. 9 is again given in Fig. 23. The t-test probabilities for both comparisons are too large to permit an unequivocal rejection of the null hypothesis in this case.

To summarize: document abstracts are more effective for content analysis purposes than document titles alone; further improvements appear possible when abstracts are replaced by large text portions; however, the increase in effectiveness is not large enough to reach the unequivocal conclusion that full text processing is always superior to abstract processing.

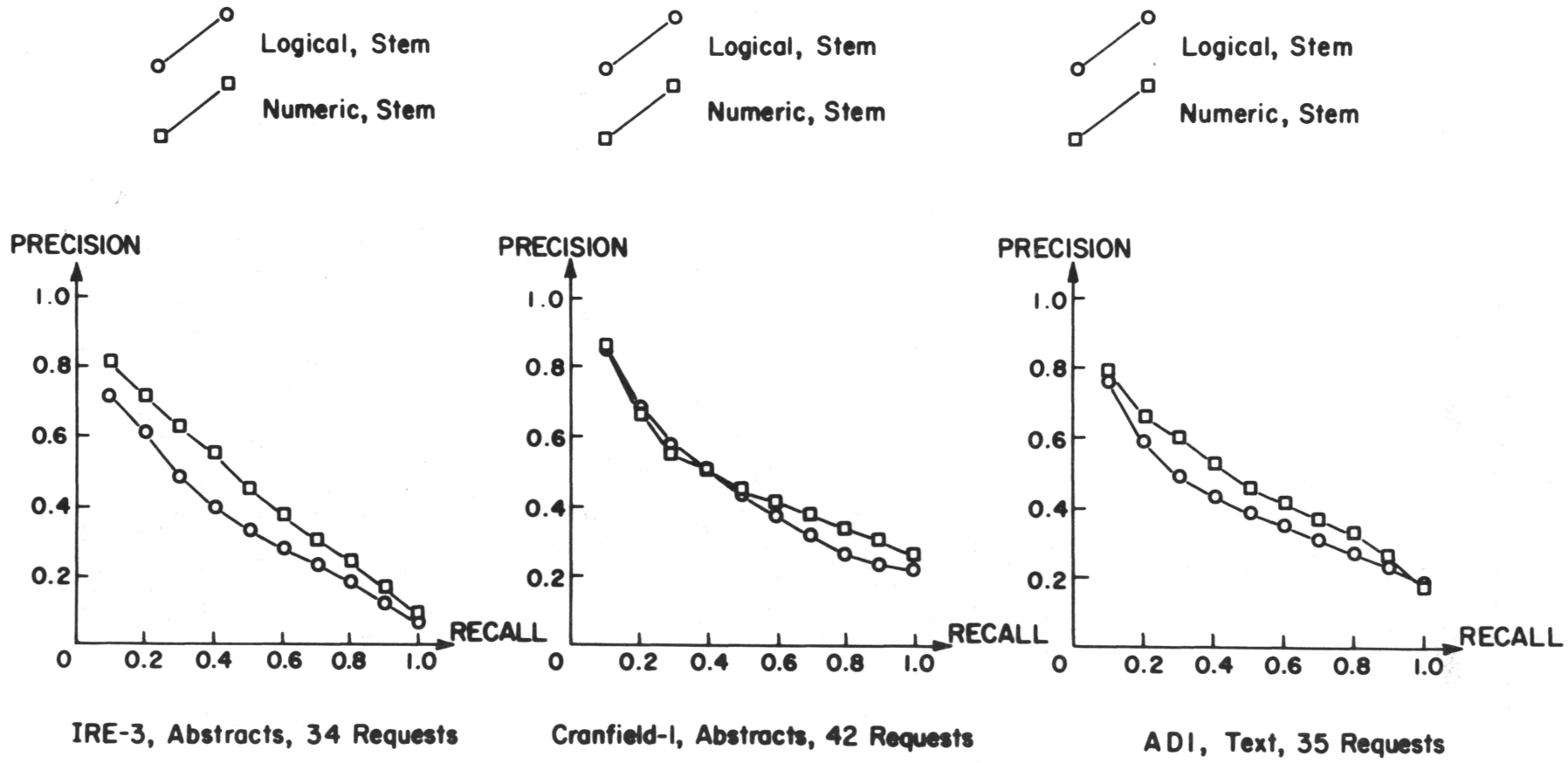
C) Matching Functions and Term Weights

It is easy in an automatic text processing environment to differentiate among individual content indicators by assigning weights to the indicators in proportion to their presumed importance. Such weights can be derived in part by using the frequency of occurrence of the original text words which give rise to the various indicators, and in part as a function of the various dictionary mapping procedures. Thus, ambiguous terms which in a synonym dictionary would normally correspond to many different thesaurus classes, can be weighted less than unambiguous terms. The SMART system includes procedures for testing the effectiveness of such weighted (numeric) content indicators compared with non-weighted (logical) indicators, where all term weights are either 1 or 0 (1 if a given term is assigned to a given document and 0 if it is not).

The recall-precision graphs for the three collections previously used are shown in Fig. 10, and the corresponding significance output is reproduced in Fig. 23. In each case, weighted word stems extracted from document abstracts or full text are compared with nonweighted (logical) stems. The results are clearly in favor of the weighted process for all three collections, the largest performance differences being registered for the IRE collection in documentation. The recall-precision graph for the ADI collection also appears to show a considerable advantage for the weighted process, and this is reflected in the t-test probability of 0.0040. However, when the magnitudes of the results are disregarded, it is found that nearly as many evaluation parameters favor the nonweighted process as the weighted one for the documentation collection. The test results are therefore not wholly significant for that collection.

On the whole, it appears that weighted content indicators produce better retrieval results than nonweighted ones, and that binary term vectors should therefore be used only if no weighting system appears readily available.

Another variable affecting retrieval performance which can be easily incorporated into an automatic information system is the correlation coefficient used to determine the similarity between an analyzed search request and the analyzed documents. Two of the correlation measures which have been included in the SMART system, are the cosine and overlap correlations, which are defined as follows:



Comparison of Weighted (Numeric) and Unweighted (Logical)

Concepts using the Stem Dictionary and Cosine Correlation

Fig. 10

$$\cos(\underline{q},\underline{d}) = \frac{\sum_{i=1}^{n} \underline{d}_i \, \underline{q}_i}{\sqrt{\sum_{i=1}^{n} (\underline{d}_i)^2 - \sum_{i=1}^{n} (\underline{q}_i)^2}} \quad ,$$

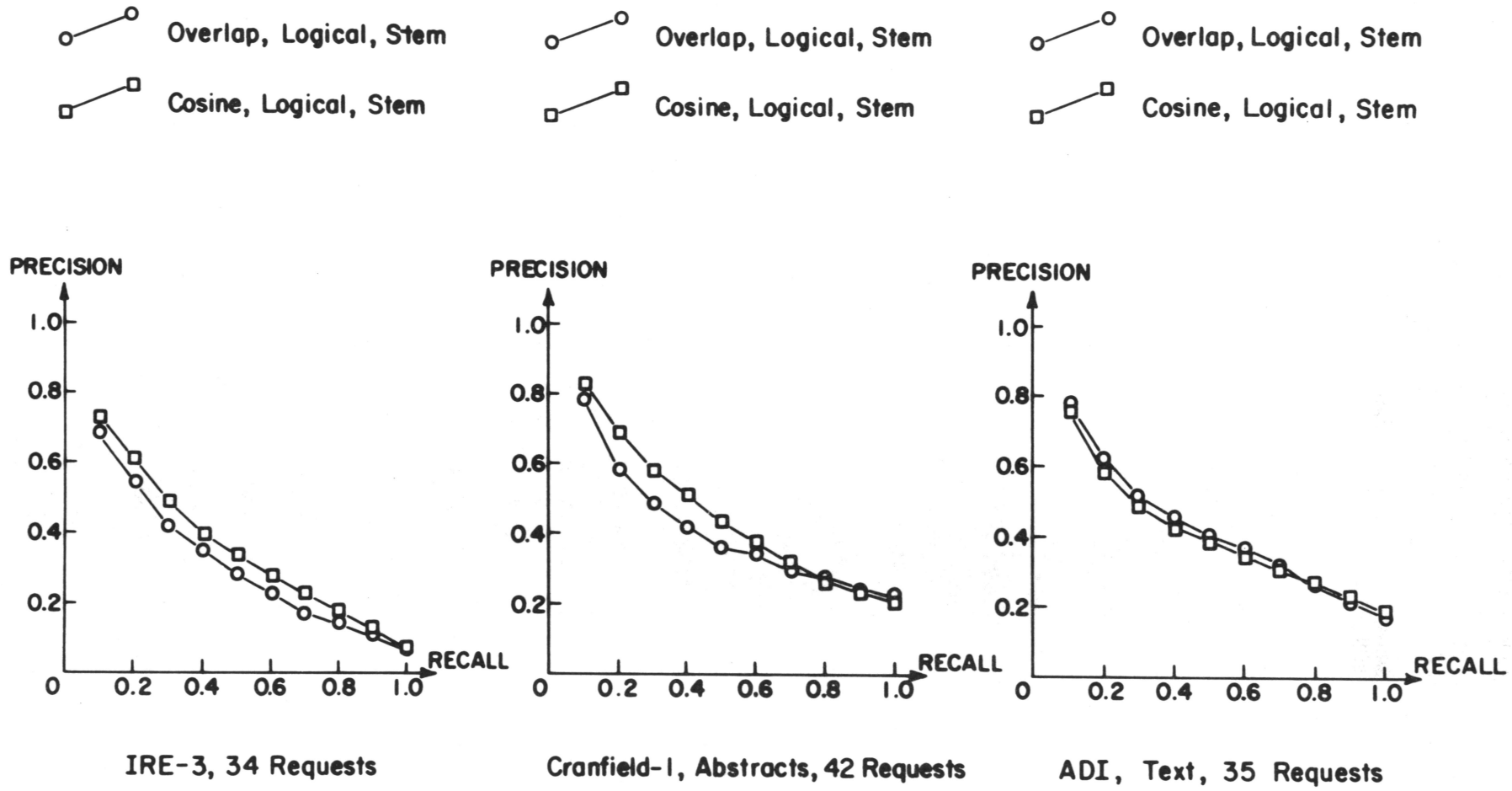
and

$$\text{ovlap}\ (\underline{q},\underline{d}) = \frac{\sum_{i=1}^{n} \min\ (\underline{q},\underline{d}_i)}{\min\left(\sum_{i=1}^{n} \underline{q}_i, \sum_{i=1}^{n} \underline{d}_i \right)} \quad ,$$

where $\underline{q}$ and $\underline{d}$ are considered to be n-dimensional vectors of terms representing an analyzed query $\underline{q}$ and an analyzed document $\underline{d}$, respectively, in a space of n terms assignable as information identifiers.

Both the cosine and the overlap functions range from 0 for no match to 1 for perfect identity between the respective vectors. The cosine correlation is more sensitive to document length, that is to the number of assigned terms, because of the factor in the denominator, and tends to produce greater variations in the correlations than the overlap measure.

A comparison of cosine and overlap matching functions is shown in the output of Fig. 11. In each case, logical (nonweighted) vectors are used with either of the two correlation methods. The results are clearly in favor of the cosine matching function, although the sign test result for the ADI collection is not sufficiently one-sided to reach a hard conclusion in that case.



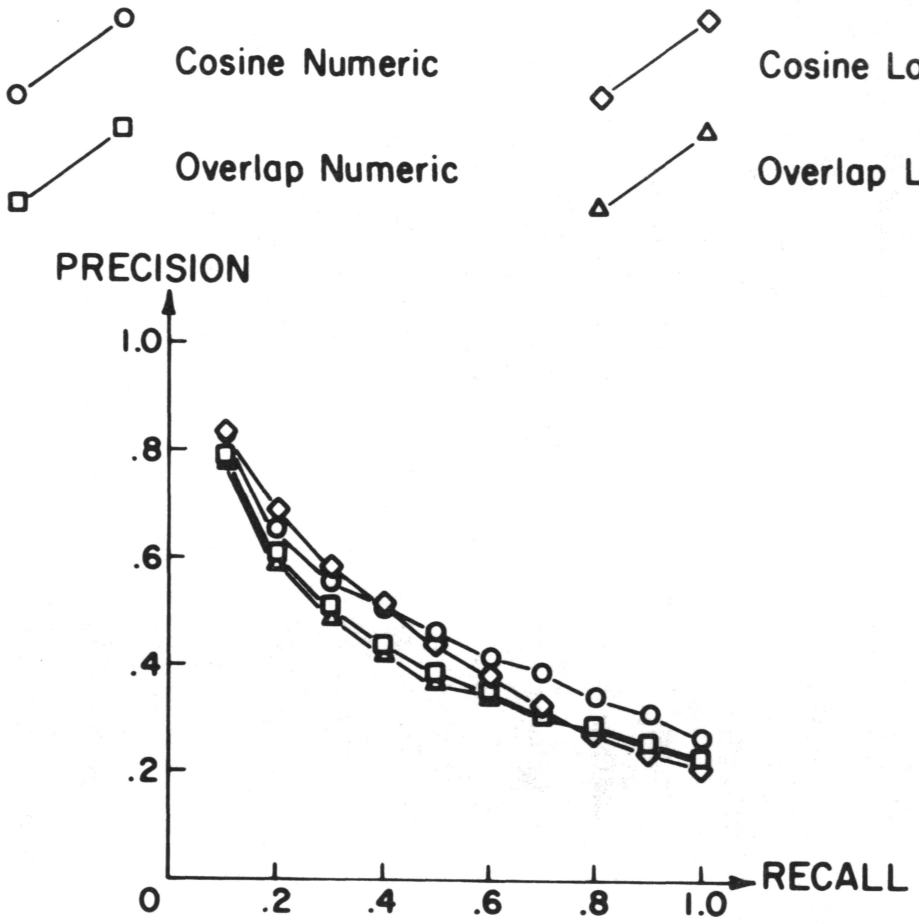
Comparison of Overlap and Cosine Matching Functions

Fig. 11

The combined effect of parameter adjustments in both the weighting and the correlation method can be studied using the output of Fig. 12, obtained for the Cranfield aerodynamics collection. The weakest method appears to be the combination of logical vectors with the overlap correlation, and the most satisfactory results are obtained with the numeric (weighted) term vectors and the cosine correlation. This confirms the results previously derived from Figs. 10 and 11.

It should be noted that the overlap-logical process corresponds to the standard keyword matching method used in almost all operational, semi-mechanized retrieval situations. In such cases, nonweighted keywords assigned to each document are compared with keywords attached to the search requests and a count is taken of the number of overlapping keywords; the resulting coefficient is then equivalent to a nonnormalized overlap function. It would appear from the results of Figs. 10 to 12, that standard keyword matching systems can be improved by the simple device of using a better matching function and assigning weights to the keywords.

The significance output corresponding to Fig. 12 is shown in Fig. 23. By scanning the columns from top to bottom, it is seen that the weighted overlap function is effectively equivalent to the nonweighted overlap function. Both of these procedures are inferior to the nonweighted cosine function, and that one is in turn inferior to the weighted cosine correlation; in the last two cases the evaluation is fully significant, and the equivalence probabilities are smaller than 1.10^{-4}.



CRANFIELD-I, 42 Requests,
Abstracts, Stems

Comparison of Correlation Coefficients and Weights

Fig. 12

To summarize: weighted content identifiers are more effective for content description than nonweighted ones, and the cosine correlation function is more useful as a measure of document-request similarity than the overlap function; advantage can therefore be taken of the computational facilities incorporated into many mechanized information systems, and service can be improved by using more sophisticated request-document matching methods.

D) Language Normalization — The Suffix Process

If natural language texts are to form the basis for an automatic assignment of information identifiers to documents, then the question of language normalization is of primary concern. Indeed, there do not then exist human intermediaries who could resolve some of the ambiguities inherent in the natural language itself, or some of the inconsistencies introduced into written texts by the authors or writers responsible for the preparation of the documents.

A large number of experiments have therefore been conducted with the SMART system, using a variety of dictionaries for purposes of language normalization in each of the three subject fields under study. The performance of the following dictionaries is studied in particular:

a) suffix "s" process, where words differing by the addition of a terminal "s" are recognized as equivalent (for example, the words "apple" and "apples" are assigned a common identifier, but not words "analyzer" and "analyzing";

b) the word stem dictionary, where all words which exhibit a common word stem are treated as equivalent; for example, "analysis", "analyzer", "analyst", and so on;

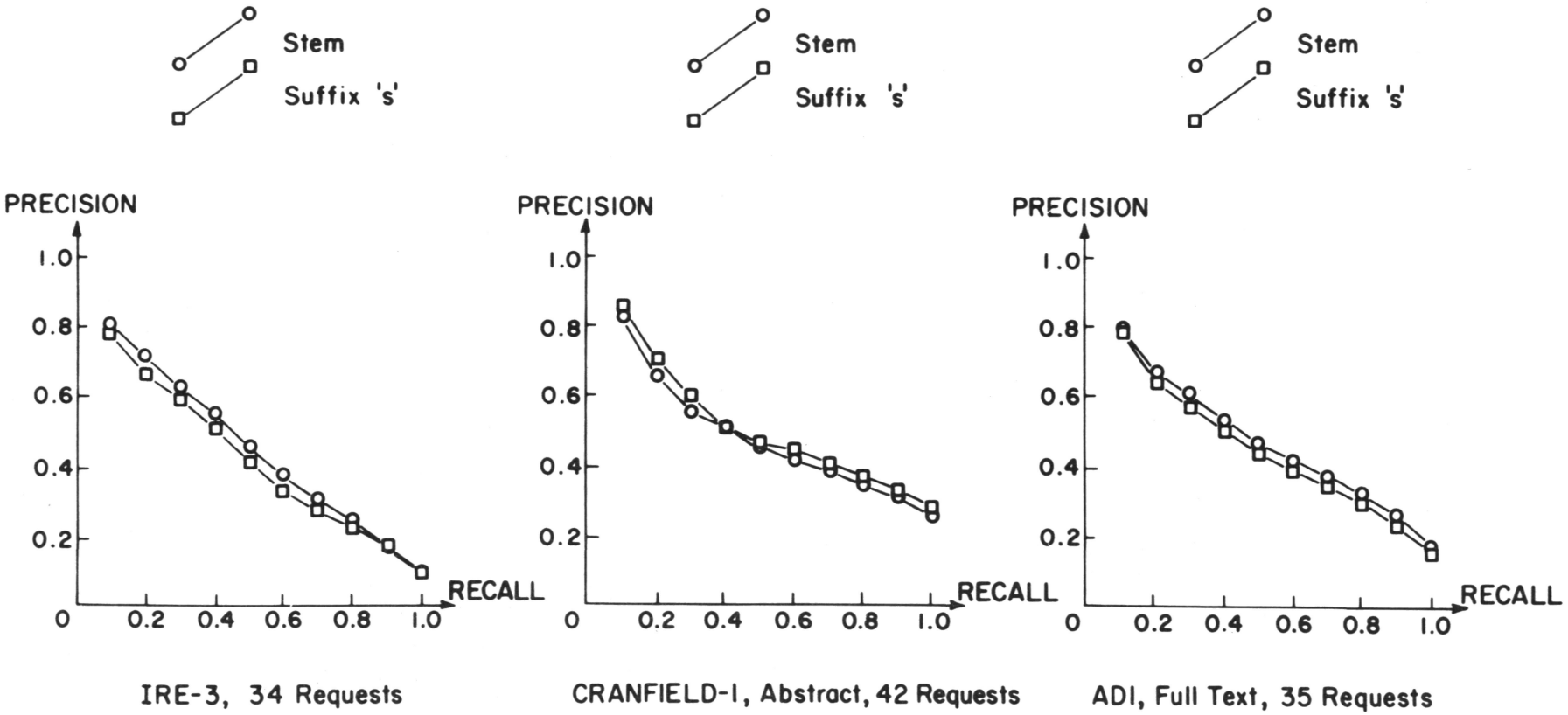
c) the synonym dictionary, or thesaurus, where a set of synonymous, or closely related terms are all placed into a common thesaurus class, thus ensuring that common identifiers are derived from all such terms;

d) the statistical phrase dictionary which makes it possible to recognize "phrases" consisting of the juxtaposition of several distinct concepts; thus if a given document contains the notion of "program", as well as the notion of "language", it could be tagged with the phrase "programming language"; the statistical phrase dictionary incorporated into the SMART system is manually constructed and contains a large variety of common noun phrases for each of the subject areas covered;

e) the concept association method in which concepts are grouped not by reference to a preconstructed dictionary, but by using statistical co-occurrence characteristics of the vocabulary under investigation.

A comparison of the "suffix 's'" dictionary with a complete "word stem" dictionary is shown in the output of Fig. 13. In the former case, the texts of documents and search requests are looked up in a table of common words so as to delete function words and other text items not of immediate interest for content analysis purposes; the final 's' endings are then deleted so as to confound words which differ only by a final 's'. In the latter case, a complete suffix dictionary is also consulted, and the original words are reduced to word stem form before request identifiers are matched with document identifiers.

The results obtained from the experiments represented in Fig. 13 are contradictory, in the sense that for two of the collections used (IRE-3 and ADI) the more thorough normalization inherent in the word stem process, compared with suffix 's' recognition alone, improves the search effectiveness;



IRE-3, 34 Requests

CRANFIELD-1, Abstract, 42 Requests

ADI, Full Text, 35 Requests

Evaluation of Suffix Cut-off Process

Fig. 13

for the third collection (Cranfield), the reverse result appears to hold. For none of the collections is the improvement of one method over the other really dramatic, so that in practice either procedure might reasonably be used.

The discrepancy between the IRE and ADI results, on the one hand, and the Cranfield results, on the other, may be caused by differences in the respective vocabularies. Specifically, the Cranfield texts are substantially more technical in nature, and the collection is more homogeneous than is the case for the other collections. To be able to differentiate between the various document abstracts, it is then important to maintain finer distinctions for the Cranfield case than for ADI and IRE, and these finer differences are lost when several different words are combined into a unique class through the suffix cut-off process. The argument can be summarized by stating that dictionaries and word normalization procedures are most effective if the vocabulary is redundant and relatively nontechnical; in the reverse case, such procedures may not in fact result in processing advantages.

E) Synonym Recognition

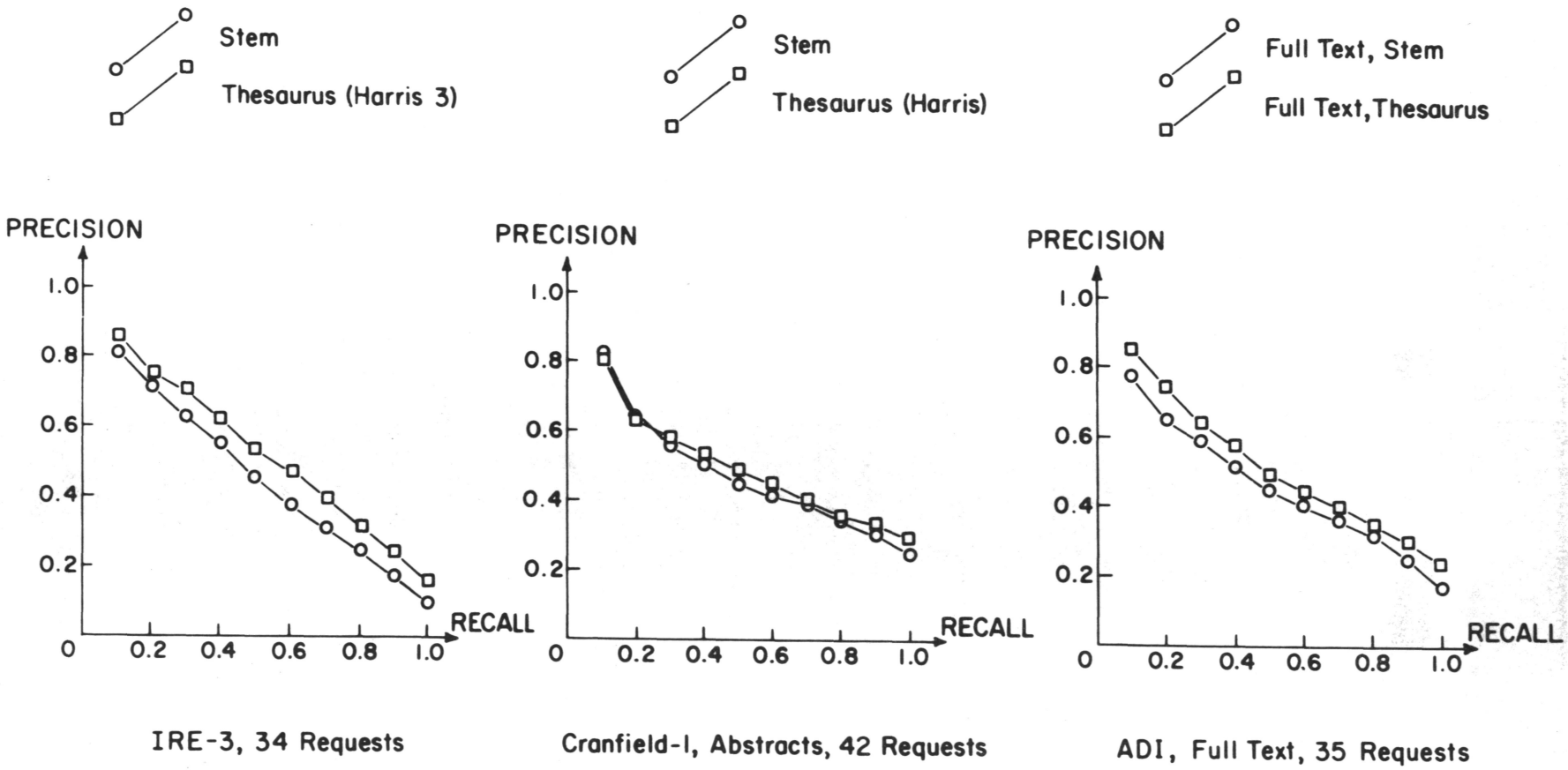
One of the perennial problems in automatic language analysis is the question of language variability among authors, and the linguistic ambiguities which result. Several experiments have therfore been performed using a variety of synonym dictionaries for each of the three subject fields under study ("Harris 2" and "Harris 3" dictionaries for the computer literature, "Quasi-synonym" or "QS" lists for aeronautical engineering, and regular thesaurus for documentation). Use of such a synonym dictionary permits the replacement of a variety of related terms by similar concept identifiers, 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. 14 which represents a comparison of the word-stem matching procedure with a process including a thesaurus look-up operation for the recognition of synonyms shows that considerable improvements in performance are obtainable by means of suitably constructed synonym dictionaries. The improvement is again smallest for the Cranfield collection in part for the reasons already stated in the last subsection, and in part because the dictionary available for this collection was not originally constructed to mesh in with the SMART retrieval programs. However, in the present case, the synonym recognition seems to benefit the Cranfield material also. The significance output for Fig. 14 shows that all thesaurus improvements are fully significant, with the exception of the t-test for the Cranfield collection. Thus only for Cranfield can the null hypothesis not be rejected unequivocally.

The differences observed in the performance of the various synonym dictionaries suggest that not all dictionaries are equally useful for the improvement of retrieval effectiveness. The experiments conducted with the SMART system in fact lead to the following principles of dictionary construction [17]:

a) very rare terms which occur in a representative sample document collection with insufficient frequency should not be placed into separate categories in the dictionary, but should be combined if possible with other rare terms to form large classes, since low frequency categories provide few matches between stored items and the search requests;



Comparison of Synonym Recognition (Thesaurus) with

Word Stem Matching Process

Fig. 14

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

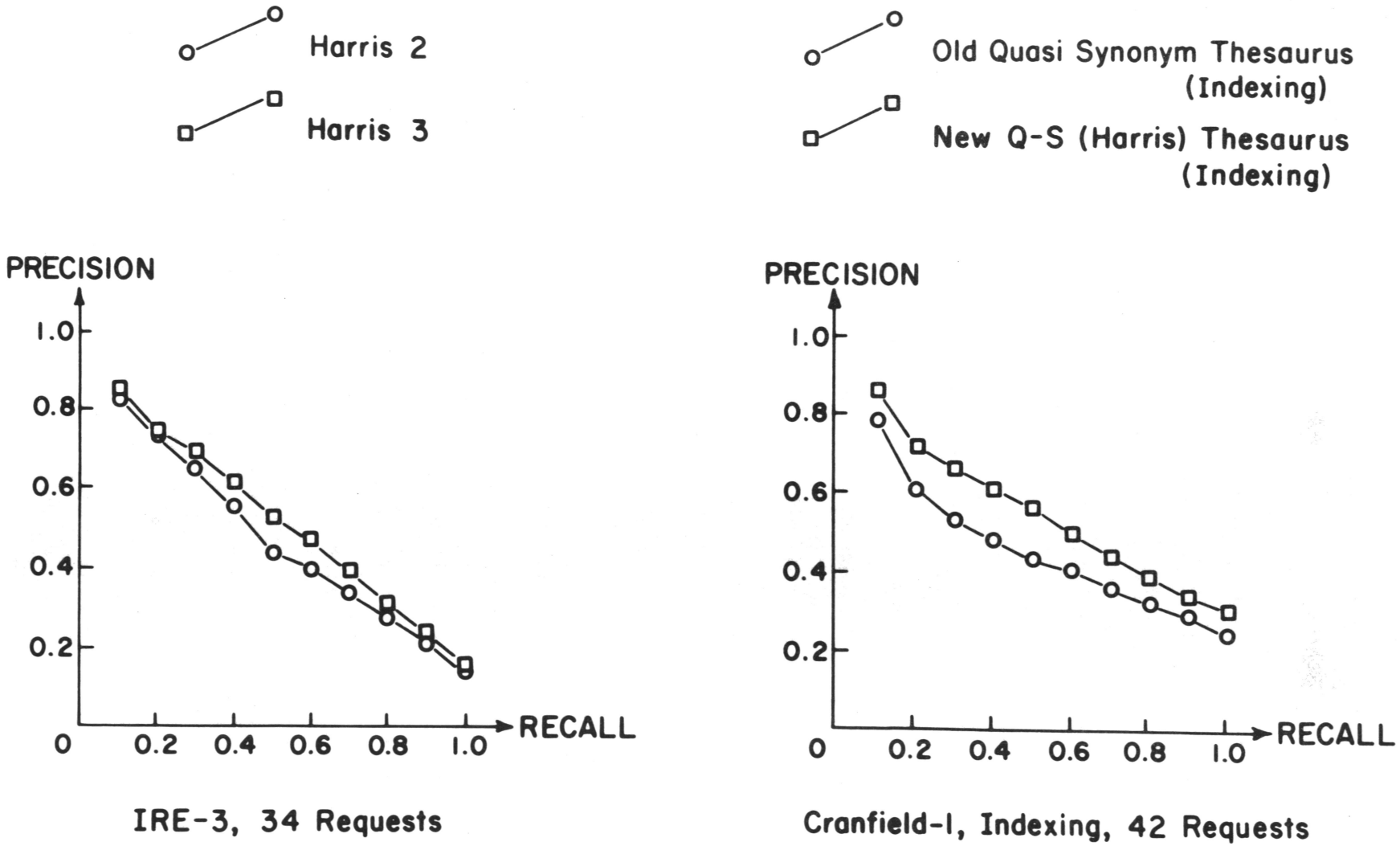
c) terms which have no special significance in a given technical subject area (such as "begin", "indicate", "system", "automatic", etc.) should not be included;

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 related 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 into semiautomatic programs for the construction of synonym dictionaries, using word frequency lists and concordances derived from a representative sample document collection. [17]

The differences in search effectiveness for two sets of two synonym dictionaries are shown in Fig. 15. The less effective dictionaries (Harris-2 for the IRE collection, and "Old Quasi-Synonym" for the Cranfield) were in each case manually constructed by specialists using ad hoc procedures set up for the occasion. The other two dictionaries are improved versions obtained manually be using some of the dictionary construction principles previously listed. The significance output shows that fully significant improvements are obtained from one dictionary version to the next. It may be noted in this connection that the synonym recognition results of the main Cranfield experiments [18] were obtained with the "old" less effective synonym dictionary, rather than with the new one.



Comparison of Thesaurus Dictionaries

Fig. 15

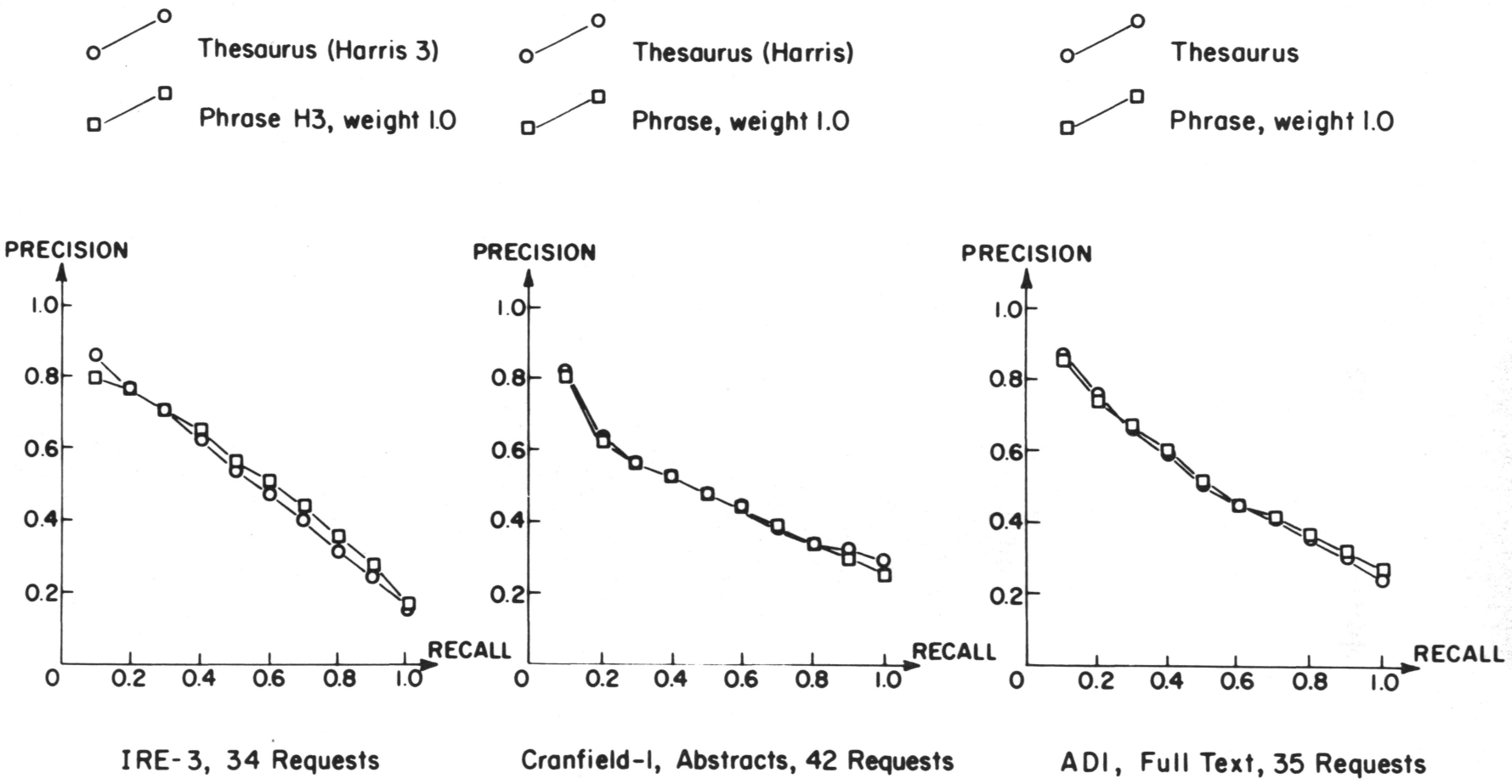
To summarize: it appears that dictionaries providing synonym recognition produce statistically significant improvements in retrieval effectiveness compared with the word stem matching process; the improvement is largest for dictionaries obeying certain principles with regard to the word groupings which are incorporated.

F) Phrase Recognition

The SMART system makes provision for the recognition of "phrases" to identify documents and search requests, rather than only individual concepts alone. 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. [19]

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.

Fig. 16 shows the results of the phrase look-up procedure compared with the equivalent process using only a synonym dictionary. It



Comparison of Phrase Dictionaries

Fig. 16

is seen that for two of the collections the phrase dictionary offers improvements in certain ranges of the recall and precision curve. The output of Fig. 23 indicates, however, that the improvements are not significant, and on the whole the phrase dictionary does not appear to offer any real help in the middle recall range. Whether this result is due to the recognition of false phrases where phrase components are all present in the text but do not really belong together (such as in the phrase "solid state" in the sentence "people whose knowledge is solid state that computer processing is efficient") remains to be seen. The evidence available would seem to indicate that the presence of such false phrases is quite rare, and that a more serious deficiency is the small size of the statistical phrase dictionary from which many potentially useful phrases may be absent.

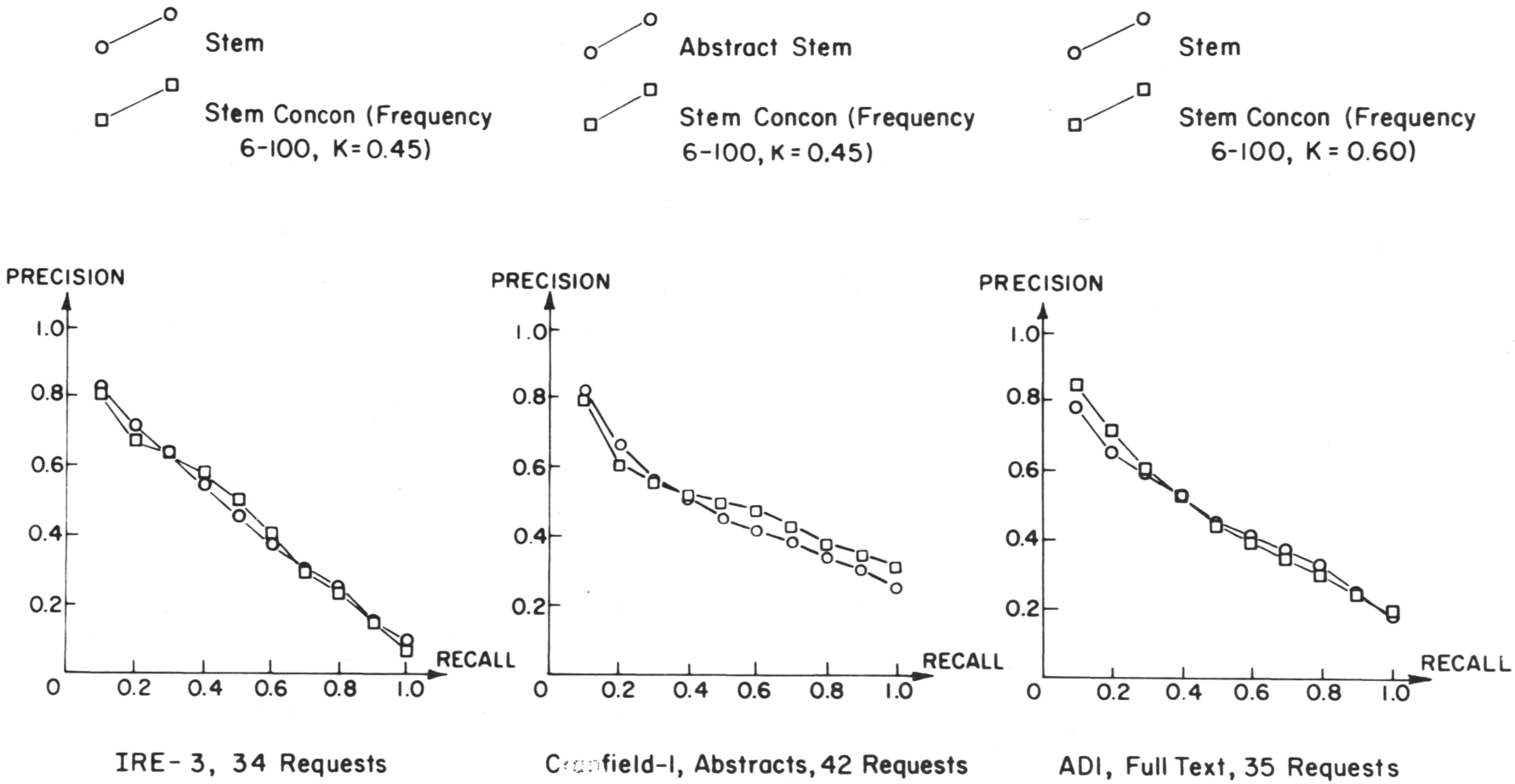
Phrases can also be recognized by using not a preconstructed dictionary of phrase components, but the statistical properties of the words in a text. Specifically, if two given terms co-occur in many of the documents of a collection, or in many sentences within a given document, a non-zero correlation 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 have more general validity. [20,21]

Many possible methods exist for the generation of statistical word associations. Specifically, by suitably varying several parameters, a number of different types of term associations can be recognized; further-

more, once an association between term pairs is introduced, it is possible to assign to it a smaller or a greater weight. Two main parameters that can be used in this connection are the cut-off value K in the association coefficient below which an association between terms 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 may not be observable under any stricter conditions. Increasingly more restrictive association procedures, applied first only to words 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.

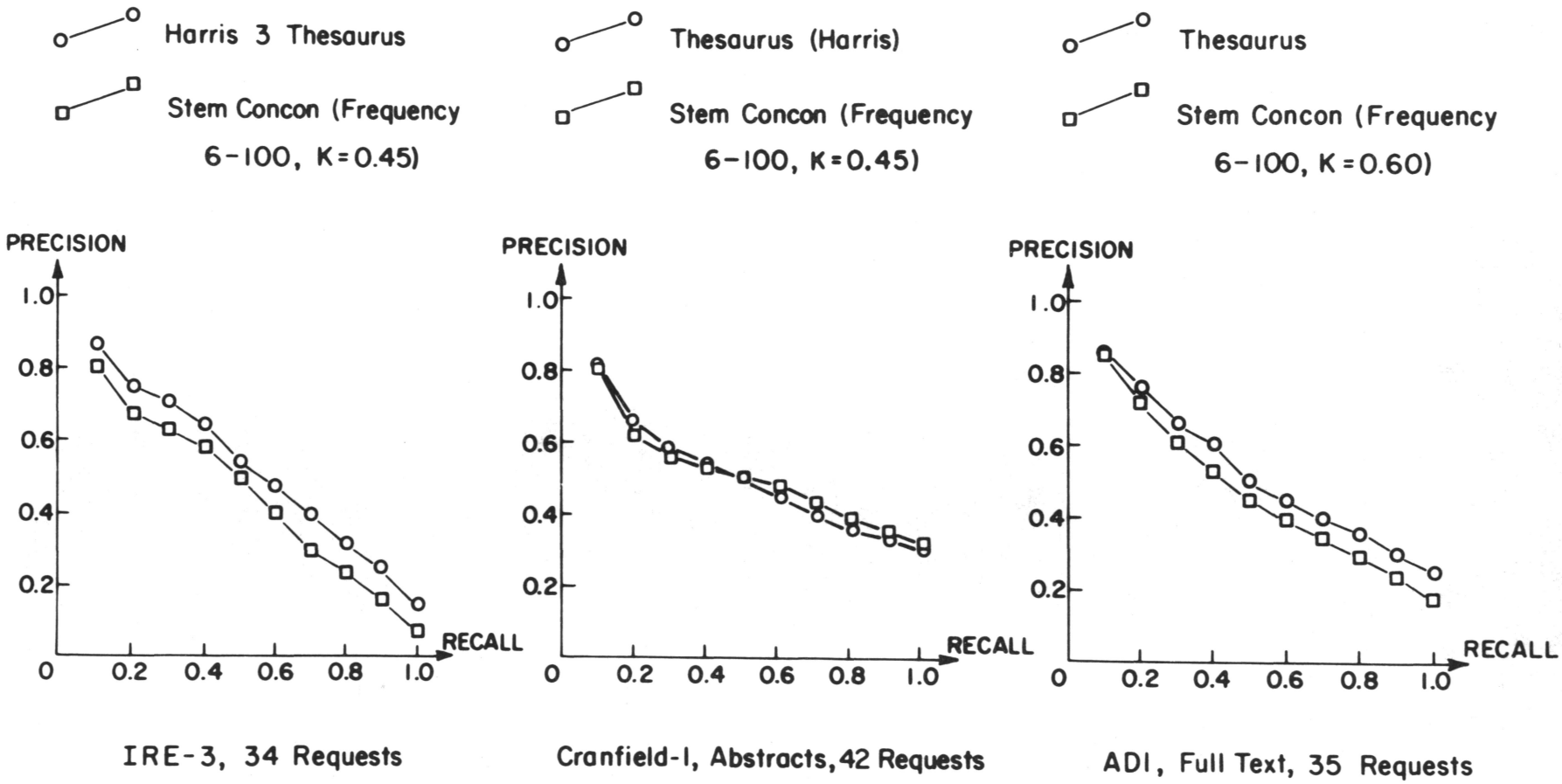
Fig. 17 shows a comparison of the word stem matching process with the statistical term-term association method (labelled "stem con-con" in Fig. 17 to indicate a concept-concept association in which word stems are manipulated). The applicable frequency restrictions for the concept pairs and the cut-off values K are also included in Fig. 17. The output of Fig. 17 and the corresponding significance computations indicate that for the Cranfield collections in particular, the term associations provide some improvement over the word stem process; local improvements for certain recall ranges are also noticeable for the ADI and IRE collections. Only for Cranfield does the sign test appear to be of some statistical significance, so that based on the present tests, no strong claims of overall effectiveness can be made for the association process.

This conclusion is reinforced with the output of Fig. 18 and the corresponding significance calculations. Here a comparison is made between the thesaurus look-up process and the word stem association method. This time, the



Comparison of word stem dictionary with addition of Statistical Word-Word Association (Stem Concon).

Fig. 17



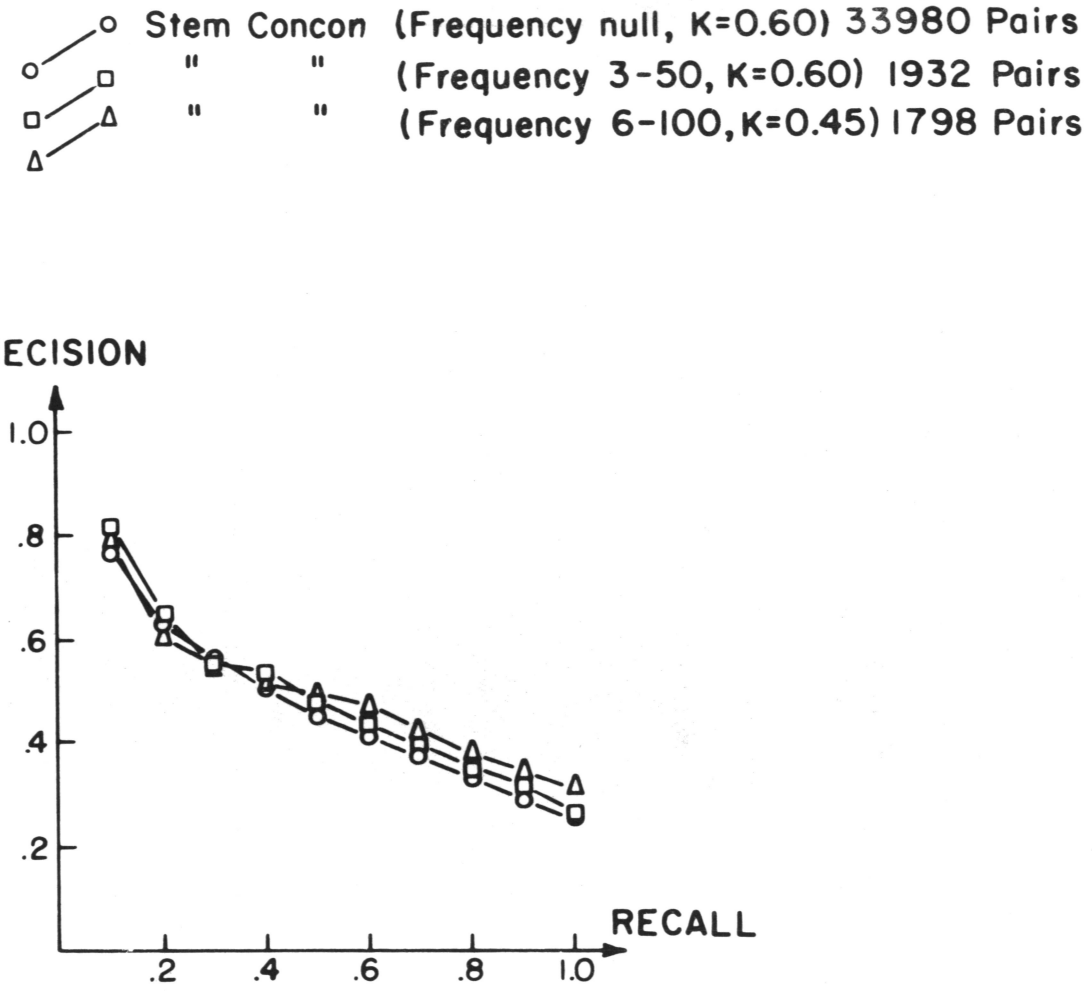
Comparison of Thesaurus Performance with Statistical Word-Word Association (Stem Concon).

Fig. 18

advantage is clearly and significantly with the more powerful thesaurus method for both the ADI and IRE collections. For Cranfield, the advantage is still slightly with the word stem association process particularly at the high recall end, where the more exhaustive indexing procedure represented by the stem associations supplies additional useful information identifiers, and serves therefore to maintain the finer distinctions among the Cranfield documents. However, the superiority of the association method is not statistically significant for Cranfield, so that the conclusion previously reached must stand.

Fig. 19 shows a comparison of various word stem association strategies performed for the Cranfield collection. The output suggests that the more restrictive association processes are more effective as a retrieval aid than the more general ones. Specifically, as the number of generated association pairs grows, too many of them appear to become spurious, thus depressing the retrieval performance. The sign-test output of Fig. 23, corresponding to the graphs of Fig. 19, shows that the differences in performance between the various association methods seem of some significance, so that in practice, limitations should be imposed on the number and types of associated terms actually used.

To summarize: the phrase generation methods, whether implemented by dictionary look-up or by statistical association processes, appear to offer improvements in retrieval effectiveness for some recall levels by introducing new associated information identifiers not originally available; the improvement is not, however, sufficiently general or substantial, when averages over many search requests are considered, to warrant incorporation into automatic information systems, except



Comparison of Word-word Association Strategies (Stem Concord)

Fig. 19

under special circumstances where suitable control procedures can be maintained.

G) Hierarchical Expansion

Hierarchical arrangements of subject identifiers are used in many standard library classification systems, and are also incorporated into many non-conventional information systems. Subject hierarchies are useful for the representation of generic inclusion relations between terms, and they also serve to broaden, or narrow, or otherwise "expand", a given content description by adding hierarchically related terms to those originally available. Specifically, given an entry point in the hierarchy, it is possible to find more general terms by going "up" in the hierarchy (expansion by parents), and more specific ones by going "down" (expansion by sons); related terms which have the same parent can also be obtained (expansion by brothers), and finally any available cross-references between individual entries can be identified (expansion by cross-references).

A hierarchical arrangement of thesaurus entries was used with the IRE collection to evaluate the effectiveness of the hierarchical expansion procedures. In the test process carried out with the SMART system, the concepts present in both the document and request vectors were looked up in the hierarchy and appropriately related hierarchy entries were added to the original content identifiers. The expanded document vectors which resulted were then matched with the requests, and documents were arranged in decreasing correlation order as usual.

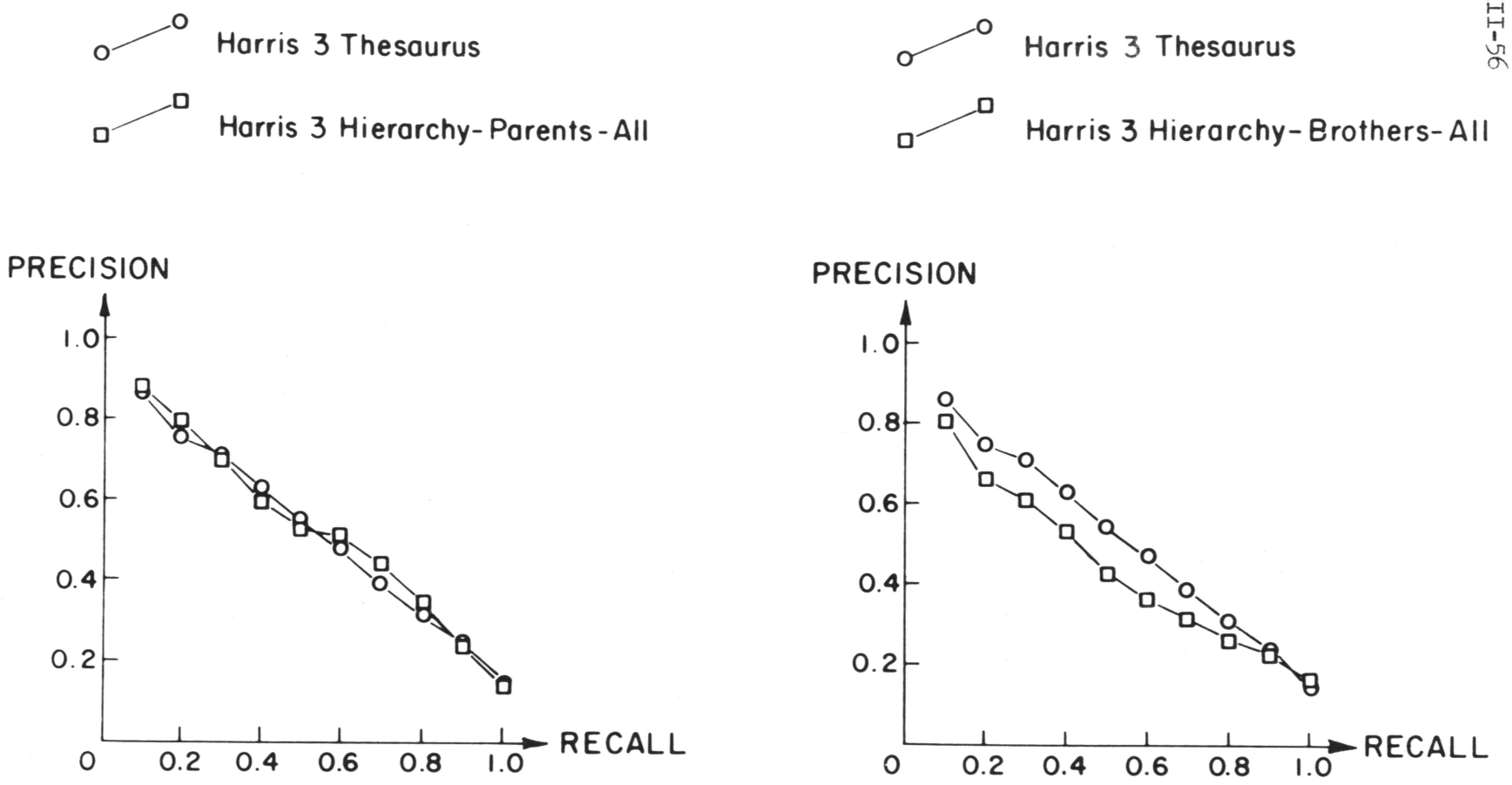
A comparison of the standard thesaurus process with the four hierarchical expansions previously described is shown in Figs. 20 and 21, respectively, and the corresponding significance output is included in Fig. 23.

It is seen that in each case the standard thesaurus process alone is superior; moreover, the equality hypothesis can be rejected unequivocally for the expansions by brothers, sons, and cross-references. A question exists only for the expansion by parents, where more general terms are added to the original identifiers. Fig. 20 (a) shows, in particular, that this expansion process does in fact improve retrieval performance for certain recall levels.

There exist of course many alternative methods for using a hierarchy. It is possible, for example, to expand requests without expanding the documents, or vice-versa; terms obtained from the hierarchy can also replace the original content identifiers instead of being added to them. In general, the expansions tend to produce 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 immediate incorporation in an analysis system for automatic document retrieval.

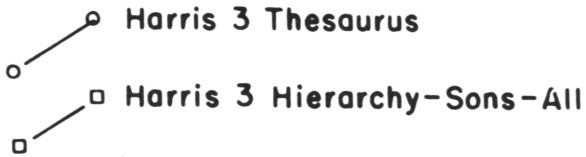
H) Manual Indexing

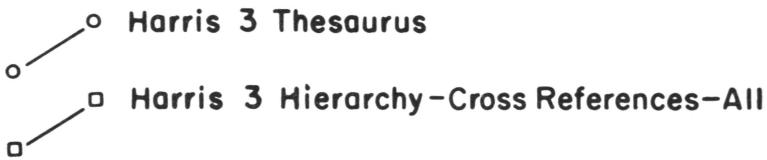
The Cranfield collection was available for purposes of experimentation both in the form of abstracts and in the form of manually assigned index terms.

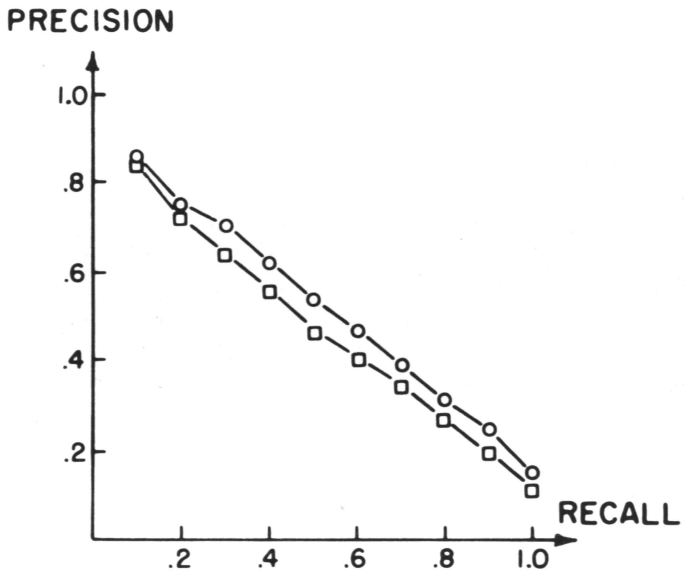


Sample Hierarchy Procedures (IRE-3, 34 Requests)

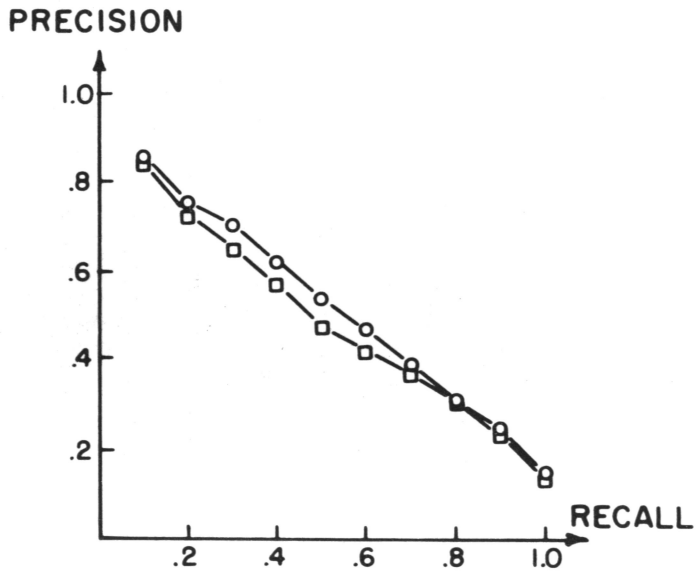
Fig. 20







a) Hierarchy Expansion by Sons



b) Hierarchy Expansion by Cross References

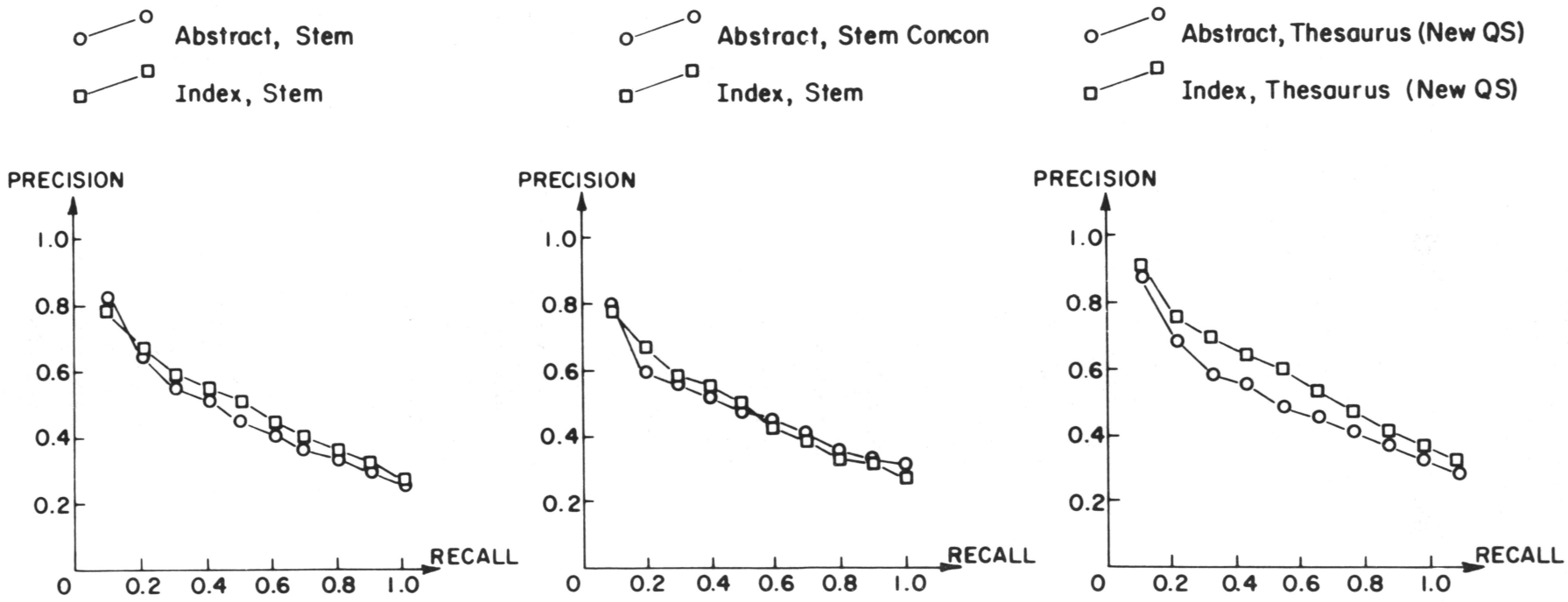
Sample Hierarchy Procedures (IRE-3, 34 Requests)

Fig. 21

The indexing performed by trained indexers is extremely detailed, consisting of an average of over 30 terms per document. As such, the indexing performance may be expected to be superior to the subject indexing normally used for large document collections. A meaningful comparison with standard manual keyword indexing systems may therefore not be possible. A comparison of index term performance with certain automatic procedures using document abstracts is represented in Fig. 22, together with the corresponding significance output in Fig. 23. Figs. 22 (a) and 22 (b) show that the overall performance of a straight index term match is only slightly superior to a match of word stems abstracted from the document abstracts; for certain recall ranges, the automatic word-word association method in fact proves to be more effective than a manual index term match. In any case, Fig. 23 shows that the null hypothesis, postulating equivalence, cannot be rejected in that instance.

When the index terms are looked up in the thesaurus and a comparison is made with the thesaurus process for the document abstracts, a clearer advantage is apparent for the indexing; that is, the identification of synonyms and related terms, inherent in the thesaurus process, seems of greater benefit to the indexing than to the automatic abstract process. Even there, however, the advantage for the index term process is not fully significant.

Based on those results, it is therefore not possible to say that the automatic text processing is substantially inferior to the manual indexing method; indeed, one is tempted to say that the efforts of the trained indexers may well have been superflous for the collection at hand,



a) Manual Indexing vs Abstract, (Word Stem Process)

b) Manual Indexing (Stem) vs Abstract (Word-Word Association)

c) Manual Indexing vs Abstract, Thesaurus

Comparison of Manual Indexing with Text Processing

(Cranfield-I, 42 Requests)

Fig. 22

since equally effective results could be obtained by simple word matching techniques. Such a result appears even more probable in the case of larger or less homogeneous collections, where the manual indexing tends to be less effective 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 in such cases does not necessarily decay as the collections grow larger, and the evaluation output may then be more favorable for the automatic procedures.

4. Concluding Comments

A summary of the main evaluation output is contained in Fig. 24, where eight processing methods are presented in order for the three document collections used. The measure used to rank the output is a combined coefficient consisting of the sum of the normalized recall and the normalized precision. The following principal conclusions can be drawn from the data of Fig. 24:

a) the order of merit for the eight methods is generally the same for all three collections, with the possible exception of the suffix 's' method which performs better than average for CRAN-1, and worse than average for ADI;

b) the performance range of the methods used is smaller for the Cranfield collection than for the other two collections;

c) the use of logical vectors (disregarding term weight), overlap correlation, and titles only is always less effective than the use of weighted terms, cosine correlation, and full document abstracts;

d) the thesaurus process involving synonym recognition always performs more effectively than the word stem or suffix 's' methods when synonyms are not recognized;

e) the thesaurus and statistical phrase methods are substantially equivalent; other dictionaries perform less well (with the exception of suffix 's' for Cranfield).

These results indicate that in automatic systems weighted terms should be used, derived from document excerpts whose length is at least equivalent to that of an abstract; furthermore, synonym dictionaries should be incorporated wherever available. Other, local improvements may be obtainable by incorporating phrases, hierarchies, and word-word association techniques. The Cranfield output shows that the better automatic text processing methods (abstracts - thesaurus) may not be substantially inferior to the performance obtained with manually assigned index terms.

A comparison of the test results obtained here with other related studies is difficult to perform. For the most part, only fragmentary results exist which do not lend themselves to a full analysis. [16,22] The Cranfield project studies contain the only available extensive test results, including the performance of manually assigned index terms, phrases, and dictionary concepts together with a wide variety of "recall devices" (procedures that broaden or generalize the meaning of the terms), and "precision devices" (procedures that add discrimination and narrow the coverage of the terms). [18] The principal conclusions reached by the Cranfield project are also borne out by the SMART studies: that phrase languages are not substantially superior to single terms as indexing devices, that synonym dictionaries improve performance, but that other dictionary types, such as hierarchies are not as effective as expected.

Future experiments leading to the design of automatic information systems should be performed in different subject areas with larger document collections.

Retrieval Methods being Compared	Corresponding Graph Number	Document Collection					
		IRE-3		CRAN-1		ADI	
		P_s	P_t	P_s	P_t	P_s	P_t
A. Title Stem B. Abstract Stem	Fig. 8	0.0000 (B>A)	0.0000 (B>A)	0.0000 (B>A)	0.0000 (B>A)	0.0000 (B>A)	0.0000 (B>A)
A. Abstract Stem B. Full Text Stem	Fig. 9	-	-	-	-	0.1420 (B>A)	0.0892 (B>A)
A. Abstract Thesaurus B. Full Text Thesaurus	Fig. 9	-	-	-	-	0.0064 (B>A)	0.0987 (B>A)
A. Numeric Stem B. Logical Stem	Fig. 10 Fig. 12	0.0000 (A>B)	0.0000 (A>B)	0.0000 (A>B)	0.0000 (A>B)	0.3736 (A>B)	0.0040 (A>B)
A. Cosine Logical Stem B. Overlap Logical Stems	Fig. 11 Fig. 12	0.0000 (A>B)	0.0000 (A>B)	0.0000 (A>B)	0.0000 (A>B)	0.0891 (A>B)	0.0148 (A>B)
A. Overlap Numeric Stem B. Overlap Logical Stem	Fig. 12	-	-	0.3497 (B>A)	0.1427 (B>A)	-	-
A. Overlap Numeric Stem B. Cosine Numeric Stem	Fig. 12	-	-	0.0000 (B>A)	0.0000 (B>A)	-	-
A. Word Stem B. Suffix 's'	Fig. 13	0.0000 (A>B)	0.0000 (A>B)	0.0000 (B>A)	0.0000 (B>A)	0.0000 (A>B)	0.0000 (A>B)
A. Thesaurus B. Word Stem	Fig. 14	0.0000 (A>B)	0.0000 (A>B)	0.0020 (A>B)	0.1483 (A>B)	0.0000 (A>B)	0.0000 (A>B)
A. Old Thesaurus B. New Thesaurus	Fig. 15	0.0000 (B>A)	0.0000 (B>A)	0.0000 (B>A)	0.0000 (B>A)	-	-
A. Thesaurus B. Phrases, Weight 1.0	Fig. 16	1.0000 (A>B)	0.8645 (B>A)	0.0001 (A>B)	0.1120 (A>B)	0.0391 (B>A)	0.1171 (B>A)

Fig. 23

Retrieval Methods being Compared	Corresponding Graph Number	Document Collection					
		IRE-3		CRAN-1		ADI	
		P_s	P_t	P_s	P_t	P_s	P_t
A. Word Stem B. Stem Concon	Fig. 17	0.1948 (A>B)	0.0566 (A>B)	0.0086 (B>A)	0.0856 (B>A)	0.4420 (B>A)	0.6521 (B>A)
A. Concon all, 0.60 B. Concon 3-50, 0.60	Fig. 19	-	-	0.0000 (B>A)	0.0525 (B>A)	-	-
A. Concon 3-50, 0.60 B. Concon 6-100, 0.45	Fig. 19	-	-	0.0001 (B>A)	0.0991 (B>A)	-	-
A. Thesaurus B. Hierarchy Parents	Fig. 20	0.1047 (A>B)	0.1676 (A>B)	-	-	-	-
A. Thesaurus B. Hierarchy Brothers	Fig. 20	0.0000 (A>B)	0.0000 (A>B)	-	-	-	-
A. Thesaurus B. Hierarchy Sons	Fig. 21	0.0000 (A>B)	0.0000 (A>B)	-	-	-	-
A. Thesaurus B. Hierarchy Cross-Ref	Fig. 21	0.0000 (A>B)	0.0000 (A>B)	-	-	-	-
A. Index Stem B. Abstract Stem	Fig. 22	-	-	0.0465 (A>B)	0.0415 (A>B)	-	-
A. Index Stem B. Stem Concon	Fig. 22	-	-	0.0020 (A>B)	0.0176 (A>B)	-	-
A. Index New Thesaurus B. Abstract, New Thesaurus	Fig. 22	-	-	0.0019 (A>B)	0.0001 (A>B)	-	-

Fig. 23 continued

Combined Significance Output

Order	IRE-3			CRAN-1			ADI		
	D	Method	M	D	Method	M	D	Method	M
1	D4	Stat. Phrase	1.686	D3	Thesaurus	1.579	D4	Stat. Phrase	1.456
2	D3	Thesaurus	1.665	D1	Suffix 's'	1.574	D3	Thesaurus	1.448
3	D2	Stems	1.570	D4	Stat. Phrase	1.566	D5	Concon	1.367
4	D5	Concon	1.559	D5	Concon	1.556	D2	Stems	1.335
5	D1	Suffix 's'	1.530	D2	Stems	1.535	D2	No Weights	1.294
6	D2	No Weights	1.494	D2	No Weights	1.477	D2	Title Only	1.293
7	D2	Overlap	1.455	D2	Title Only	1.430	D1	Suffix 's'	1.283
8	D2	Title Only	1.369	D2	Overlap	1.407	D2	Overlap	1.241
Range	0.317			0.172			0.215		

Overall Merit for Eight Processing Methods Used
with Three Document Collections

M: Merit Measure
(Normalized Recall
plus Normalized
Precision)

D: Dictionary Used
(D1 : Suffix 's'
D2 : Word Stem
D3 : Thesaurus
D4 : Stat. Phrase
D5 : Word - Word Association)

Fig. 24

Furthermore, in addition to the content analysis tests, it becomes increasingly important to evaluate also the search procedures likely to be used in an automatic systems environment, and particularly those real-time search methods where the user can control the search strategy to some extent, by providing suitable feedback information. Work in this direction is continuing. [7,23,24]

Acknowledgement: The assistance of Mr. Cyril Cleverdon and Mr. Michael Keen in making available the documents and dictionaries used by the Aslib-Cranfield Research Project is gratefully asknowledged. Mr. Keen was also instrumental in preparing many of the output graphs included in this study.

References

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

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

[3] G. S. Simpson and C. Flanagan, Information Centers and Services, in Annual Review of Information Science and Technology, C. Cuadra, editor, Chapter XII, Wiley, 1966.

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

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

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

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

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

[9] C. W. Cleverdon, The Testing of Index Language Devices, Aslib Proceedings, Vol. 5, No. 4, April 1965.

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

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

[12] G. Salton, Evaluation of Computer - based Retrieval Systems, Proceedings of the FID Congress 1965, Spartan Books 1966.

[13] R. A. Fisher, Statistical Methods for Research Workers, Hafner Publishing Co., Inc., New York, 1954.

References (contd.)

[14] J. L. Hodges and E. L. Lehmann, Basic Concepts of Probability and Statistics, Holden Day, San Francisco, 1964.

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

[16] V. E. Giuliano and P. E. Jones, Study and Test of a Methodology for Laboratory Evaluation of Message Retrieval Systems, Report ESD - TR - 66 - 405 to the Electronic Systems Division, Arthur D. Little Co., August 1966.

[17] G. Salton, Information Dissemination and Automatic Information Systems, Proceedings of the IEEE, Vol. 54, No. 12, December 1966.

[18] C. Cleverdon and M. Keen, Factors Determining the Performance of Indexing Systems, Vol. 2 - Test Results, Cranfield, 1966.

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

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

[21] 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.

[22] B. Altman, A Multiple Testing of the Natural Language Storage and Retrieval ABC Method: Preliminary Analysis and Test results, American Documentation, Vol. 18, No. 1, January 1967.

[23] E. M. Keen, Semi-Automatic User Controlled Search Strategies, Fourth Annual National Colloquium on Information Retrieval, Philadelphia, 1967.

[24] G. Salton, Search Strategy and the Optimization of Retrieval Effectiveness, Report No. ISR-12 to the National Science Foundation, Cornell University, Department of Computer Science, 1967.