IV. THE EVALUATION OF AUTOMATIC RETRIEVAL PROCEDURES - SELECTED TEST RESULTS USING THE SMART SYSTEM

G. Salton*

ABSTRACT

The generation of effective methods for the evaluation of information retrieval systems and techniques is becoming increasingly important as more and more systems are designed and implemented. The present section deals with the evaluation of a variety of automatic indexing and retrieval procedures incorporated into the SMART automatic document retrieval system.

The design of the SMART system is first briefly reviewed. The document file, search requests, and other parameters affecting the evaluation system are then examined in detail, followed by a description of the measures used to assess the effectiveness of the retrieval performance. The main test results are given and tentative conclusions are reached concerning the design of fully automatic information systems.

1. Introduction

The evaluation of information retrieval systems and of techniques for indexing, storing, searching and retrieving information has become of increasing importance in recent years. The interest in evaluation procedures

* Computation Laboratory, Harvard University, Cambridge, Massachusetts.
This study was supported by the National Science Foundation under grant GN-245.

stems from two main causes: first, more and more retrieval systems are being designed, thus raising an immediate question concerning performance and efficacy of these systems; and second, evaluation methods are of interest in themselves, in that they lead to many complicated problems in test design and performance, and in the interpretation of test results.

The present study differs from other reports on systems evaluation in that it deals with the evaluation of automatic, rather than conventional, information retrieval. More specifically, it is desired to compare the effectiveness of a large variety of fully automatic procedures for information analysis (indexing) and retrieval. Since such an evaluation must of necessity take place in an experimental situation, rather than in an operational environment, it becomes possible to eliminate from consideration such important system parameters as cost of retrieval, response time, influence of physical layout, personnel problems, and so on, and to concentrate fully on the evaluation of retrieval techniques. Furthermore, a number of human problems which complicate matters in a conventional evaluation procedure, including, for example, the difficulties due to inconsistency among indexers, or to the presence of search errors, need not be considered. Other problems, including those which have to do with the identification of information relevant to a given search request, and those concerning themselves with the interpretation of test results must of course be faced in an automatic system, just as in a conventional one.

The design of the SMART automatic document retrieval system is first briefly reviewed. The test environment is then described in detail, including in particular a description of the document file and of the search

requests used. Parameters are introduced to measure the effectiveness of the retrieval performance; these parameters are similar to the standard recall and precision measures, but do not require that a distinction be made between retrieved and nonretrieved documents. The main test results are then given, and some tentative conclusions are reached concerning the design of fully automatic retrieval systems.

2. The SMART Retrieval System

SMART is a fully automatic document retrieval system operating on the IBM 7094. Stored documents as well as search requests are processed without any prior manual analysis by one of several hundred possible methods, as specified at time of input. A content analysis of each incoming item is made, and the analyzed search requests are matched against the stored document collection. Items found to be relevant, that is, whose correlation with a given search request exceeds a specified threshold, are printed out in decreasing order of the correlation coefficients.†

The following facilities incorporated into the SMART system are of principal interest:

(a) a system for separating English words into <u>stems</u> and <u>affixes</u>, thus reducing a variety of different word occurrences with similar stems to a single specified form;

† A more detailed description of the systems organization is included in Ref. 1. Programming aspects and complete flowcharts are presented in Ref. 2.

(b) a thesaurus lookup system to replace synonymous word stems by a single thesaurus category, or concept;

(c) a hierarchical arrangement of 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 procedures to compute similarity coefficients based on co-occurrences of concepts within the sentences, of a given document, or within the documents of a given collection; association factors between documents can also be determined, as can clusters (rather than only pairs) of related documents, or related concepts;

(e) syntactic matching procedures which make it possible to recognize a large number of semantically equivalent, but syntactically quite different constructions, in such a way that the same set of concept numbers can be assigned to all such equivalent structures;

(f) statistical phrase matching methods which operate like the preceding syntactic phrase procedures, except that a syntactic analysis is not performed, and syntactic dependencies between phrase components are disregarded; and

(g) a dictionary updating system, designed to revise the five principal dictionaries included in the system (stem thesaurus, suffix dictionary, concept hierarchy, statistical phrases, and syntactic "criterion" phrases).

The operations of the system are built around a supervisory system which decodes the input instructions and arranges the processing sequence in accordance with the instructions received. At the present time, about

35 different processing options are available, in addition to a number of variable parameter settings. The latter are used to specify the correlation type which measures the similarity between documents and search requests, the cut-off value which determines the number of documents to be extracted as answers to search requests, and the thesaurus size.

The SMART systems organization makes it possible to evaluate the effectiveness of the various processing methods by comparing the outputs obtained from a variety of different processing runs. This is achieved by processing the same search requests against the same document collection several times, and making judicious changes in the analysis procedures between runs. It is this use of the SMART system as an evaluation tool which is of particular interest in the present context, and is therefore treated in more detail in the remaining parts of the present report.

3. The Test Environment

The parameters which control the testing procedures about to be described are summarized in Fig. 1. The data collection used consists of a set of 405 abstracts* of documents in the computer literature, published during 1959 in the IRE Transactions on Electronic Computers. The results reported are based on the processing of about 20 search requests, each of which is analyzed by approximately 15 different indexing procedures. The

---

*Practical considerations dictated the use of abstracts rather than full documents; the SMART system as such is not restricted to the manipulation of abstracts only.

Characteristic	Comment	Count
Number of documents in collection	Document abstracts in the computer field	405
Number of search requests		
(a) specific	0-9 relevant documents	10
(b) general	10-30 relevant documents	7
User population (requestor also makes relevance judgments)	Technical people and students	about 10
Number of indexing and search programs used	All search and indexing operations are automatic	15
Number of index terms per document	Varies greatly depending on indexing procedure and document	(average) 35
Number of relevant documents per request		
(a) specific		(average) 5
(b) general		(average) 15
Number of retrieved documents per request	No cutoff is used to separate retrieved from nonretrieved	405

Test Environment

Figure 1

search requests are somewhat arbitrarily separated into two groups called, respectively, "general" and "specific" requests, depending on whether the number of documents believed to be relevant to each request is equal to at least ten (for the general requests), or is less than ten (for the specific ones). Results are reported separately for each of these two request groups; cumulative results are also reported for the complete set of requests.

The user population responsible for the search requests consists of about ten technical people with background in the computer field. Requests are formulated without study of the document collection, and no document already included in the collection is normally used as a source for any given search request. On the other hand, in view of the experimental nature of the system, it cannot be stated unequivocally that an actual user need in fact exists which requires fulfillment.

An excerpt from the document collection, as it appears in computer storage, is reproduced in Fig. 2. It may be noted that the full abstracts are stored together with the bibliographic citations. A typical search request, dealing with the numerical solution of differential equations, is shown at the top of Fig. 3. Any search request expressed in English words is acceptable, and no particular format restrictions exist. Also shown in Fig. 3 are a set of documents found in answer to the request on differential equations by using one of the available processing methods. The documents are listed in decreasing order of the correlation coefficient with the search request; a short twelve-character identifier is shown for each document under the heading "answer," and full bibliographic citations are shown under "identification."

*TEXT 2MICRO-PROGRAMMING .

$MICRO-PROGRAMMING
$R. J. MERCER (UNIVERSITY OF CALIFORNIA)
$U.S. GOV. RES. REPTS. VOL 30 PP 71-72(A) (AUGUST 15, 1958) PB 126893

MICRO-PROGRAMMING . THE MICRO-PROGRAMMING TECHNIQUE OF DESIGNING THE CONTROL CIRCUITS OF AN ELECTRONIC DIGITAL COMPUTER TO FORMALLY INTERPRET AND EXECUTE A GIVEN SET OF MACHINE OPERATIONS AS AN EQUIVALENT SET OF SEQUENCES OF ELEMENTARY OPERATIONS THAT CAN BE EXECUTED IN ONE PULSE TIME IS DESCRIBED .

*TEXT 3THE ROLE OF LARGE MEMORIES IN SCIENTIFIC COMMUNICATIONS

$THE ROLE OF LARGE MEMORIES IN SCIENTIFIC COMMUNICATIONS
$M. M. ASTRAHAN (IBM CORP.)
$IBM J. RES. AND DEV. VOL 2 PP 310-313 (OCTOBER 1958)

THE ROLE OF LARGE MEMORIES IN SCIENTIFIC COMMUNICATIONS . THE ROLE OF LARGE MEMORIES IN SCIENTIFIC COMMUNICATIONS IS DISCUSSED . LARGE MEMORIES PROVIDE AUTOMATIC REFERENCE TO MILLIONS OF WORDS OF MACHINE-READABLE CODED INFORMATION OR TO MILLIONS OF IMAGES OF DOCUMENT PAGES . HIGHER DENSITIES OF STORAGE WILL MAKE POSSIBLE LOW-COST MEMORIES OF BILLIONS OF WORDS WITH ACCESS TO ANY PART IN A FEW SECONDS OR COMPLETE SEARCHES IN MINUTES . THESE MEMORIES WILL SERVE AS INDEXES TO THE DELUGE OF TECHNICAL LITERATURE WHEN THE PROBLEMS OF INPUT AND OF THE AUTOMATIC GENERATION OF CLASSIFICATION INFORMATION ARE SOLVED . DOCUMENT FILES WILL MAKE THE INDEXED LITERATURE RAPIDLY AVAILABLE TO THE SEARCHER . MACHINE TRANSLATION OF LANGUAGE AND RECOGNITION OF SPOKEN INFORMATION ARE TWO OTHER AREAS WHICH WILL REQUIRE FAST, LARGE MEMORIES .

Typical Document Prints

Figure 2

ANSWERS TO REQUESTS FOR DOCUMENTS ON SPECIFIED TOPICS SEPTEMBER 28, 1964 PAGE 83

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)
360SIMULATIN	0.5758	SIMULATING SECOND-ORDER EQUATIONS D. G. CHADWICK (UTAH STATE UNIV.) ELECTRONICS VOL 32 P 64 (MARCH 6, 1959)
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
392ON COMPUT	0.5508	ON COMPUTING RADIATION INTEGRALS R. C. HANSEN (HUGHES AIRCRAFT CO.), L. L. BAILIN (UNIV. OF SOUTHERN CALIFORNIA, AND R. W. RUTISHAUSER (LITTON INDUSTRIES, INC.) COMMMUN. ASSOC) FOR COMPUTING MACH. VOL 2 PP 28-31 (FEBRUARY, 1959)
386ELIMINATI	0.5483	ELIMINATION OF SPECIAL FUNCTIONS FROM DIFFERENTIAL EQUATIONS J. E. POWERS (UNIV. OF OKLAHOMA) COMMUN. ASSOC. FOR COMPTING MACH. VOL 2 PP 3-4 (MARCH, 1959)

Typical Search Request and Corresponding Answers

Figure 3

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

DOCUMENT — CONCEPT,OCCURS — PAGE 17

DOCUMENT	CONCEPT,OCCURS										
DIFFERNTL EQ	4EXACT	12	8ALGOR	12	13CALC	18	71EVAL	6	92DIGI	12	REGULAR THESAURUS
	110AUT	12	143UTI	12	176SOL	12	179STD	12	181QUA	24	
	269ELI	4	274DIF	36	356VEL	12	357YAW	4	384TEG	12	
	428STB	4	505APP	24							
1A COMPUTER	2INPUT	4	5LOCAT	12	10ALPH	12	15BASE	6	16BASC	6	
	31BIT	3	32REQU	3	41MCHO	8	47CHNG	6	53DATA	6	
	57DSCB	15	59AMNT	24	72EXEC	6	77LIST	4	83MAP	6	
	87ENBL	12	93ORDR	10	106NQU	6	107DGN	30	108LOO	12	
	110AUT	36	112OPE	6	119AUT	8	121MEM	4	130MEA	4	
	143UTI	12	146JOB	18	147SYS	12	149POG	36	158REL	12	
	162ROF	6	163EAS	12	168ORD	4	176SOL	12	178SYM	18	
	182SAV	4	187DIR	12	210OUT	4	212SIZ	12	216DOM	12	
	276GEM	18	327AST	12	332SEE	12	338MCH	8	340LET	3	
	346JET	6	350IFO	6	419GEM	6	501ORD	4	508ACT	6	
DIFFERNTL EQ	ACCUR	12	ALGORI	12	COMPUT	12	DIFFER	24	DIGIT	12	NULL THESAURUS
	EQU	24	EVALU	12	GIVE	12	INTEGR	12	METHOD	12	
	NUMER	12	ORDIN	12	PARTI	12	PROCED	12	RUNGE-	12	
	SOLUT	12	SPEED	12	STABIL	12	USE	12	VARIE	12	
1A COMPUTER	BAS	12	CHARAC	12	COMPUT	36	DESCRI	12	DESIGN	12	
	DIRECT	12	ENABLE	12	ESTIM	12	EXPLAI	12	FORM	12	
	GIVE	12	HANDLE	12	ILLUST	12	INDEPE	12	INFORM	12	
	MACHIN	24	OPER	12	ORD	12	ORIENT	12	PLANE	12	
	POS	12	POSS	12	PROBLE	36	PROGRA	36	RECOGN	12	
	SCANN	12	SIMPLE	12	SIZE	24	STORE	12	STRUCT	12	
	TECHNI	12	TOWARD	12	TRANSF	12	USING	12	WRITT	12	
DIFFERNTL EQ	4EXACT	12	8ALGOR	12	13CALC	18	71EVAL	6	92DIGI	12	STAT. PHRASE LOOK-UP
	110AUT	12	143UTI	12	176SOL	12	179STD	12	181QUA	24	
	269ELI	4	274DIF	36	356VEL	12	357YAW	4	375NUM	36	
	379DIF	72	384TEG	12	428STB	4	505APP	24			
1A COMPUTER	2INPUT	4	5LOCAT	12	10ALPH	12	14CODR	72	15BASE	6	
	16BASC	6	31BIT	3	32REQU	3	41MCHO	8	47CHNG	6	
	53DATA	6	57DSCB	15	59AMNT	24	72EXEC	6	77LIST	4	
	83MAP	6	87ENBL	12	93ORDR	10	106NQU	6	107DGN	30	
	108LOO	12	110AUT	36	112OPE	6	119AUT	8	121MEM	4	
	130MEA	4	143UTI	12	146JOB	18	147SYS	12	149POG	36	
	158REL	12	162ROF	6	163EAS	12	168ORD	4	176SOL	12	
	178SYM	18	182SAV	4	187DIR	12	200DA-	72	210OUT	4	
	212SIZ	12	216DOM	12	219POG	36	276GEM	18	292THK	36	
	302LOO	72	327AST	12	332SEE	48	338MCH	8	340LET	3	
	346JET	6	350IFO	6	419GEM	6	501ORD	4	508ACT	6	

STAT. PHRASES

Typical Indexing Products for Three Analysis Procedures

Figure 4

The average number of index terms used to identify each document is sometimes believed to be an important factor affecting retrieval performance. In the SMART system, this parameter is a difficult one to present and interpret, since the many procedures which exist for analyzing the documents and search requests generate indexing products with widely differing characteristics. A typical example is shown in Fig. 4, consisting of the index "vectors" generated by three different processing methods for the request on differential equations (short form "DIFFERNTL EQ") and for document number 1 of the collection (short form "1A COMPUTER ").

It may be seen from Fig. 4 that the number of terms identifying a document can change drastically from one method to another: for example, document number 1 is identified by 35 different word stems using the so-called "null" thesaurus; these 35 stems, however, give rise to 50 different concept numbers using a regular thesaurus, and to 55 concepts including statistical phrases. Weights assigned to concept numbers also change from method to method. The number of index terms per document shown in the summary of Fig. 1 (35) is therefore at best an indication, and does not properly reflect the true situation. Since no distinction is made in the evaluation procedure between retrieved and nonretrieved documents, the last indicator included in Fig. 1 (the number of retrieved documents per request) must also be put into the proper perspective. A discussion of this point is postponed until after the evaluation measures are introduced in the next few paragraphs.

4. Evaluation Measures

A. Recall and Precision

One of the most crucial tasks in the evaluation of retrieval systems is the choice of measures which reflect systems performance. In the present context, such a measurement must of necessity depend primarily on the system's ability to retrieve wanted information, and to reject nonwanted material, to the exclusion of operational criteria such as retrieval cost, waiting time, input preparation time, and so on. The last-mentioned factors may be of great practical importance in an operational situation, but do not enter, at least initially, into the evaluation of experimental procedures.

A large number of measures have been proposed in the past for the evaluation of retrieval performance.[3] Perhaps the best known of these are, respectively, recall and precision, recall being defined as the proportion of relevant material actually retrieved, and precision as the proportion of retrieved material actually relevant.† A system with high recall is one which rejects very little that is relevant, but may also retrieve a large proportion of irrelevant material, thereby depressing precision. High precision, on the other hand, implies that very little irrelevant information is produced, but much relevant information may be

---

† Precision has also been called "relevance," notably in the literature of the ASLIB-Cranfield project.[4]

missed at the same time, thus depressing recall. Ideally, one would of course hope both for high recall and high precision.[*]

Measures such as recall and precision are particularly attractive when it comes to evaluating automatic retrieval procedures, because a large number of extraneous factors which cause uncertainty in the evaluation of conventional (manual) systems are automatically absent. The following characteristics of the present system are particularly important in this connection:

(a) input errors in the conventional sense, due to faulty indexing or encoding, are eliminated, since all indexing operations are automatic;

(b) for the same reasons, conventional search errors arising from the absence of needed search terms are also excluded;

(c) errors cannot be introduced in any transition between original search request and final machine query, since the transition is now handled automatically and becomes indistinguishable from the main analysis operation;

(d) inconsistencies introduced by a large number of different indexers, and by the passage of time in the course of an experiment cannot arise; and

(e) the role of human memory as a disturbance in the generation of retrieval measurements is eliminated (this factor can be particularly troublesome when source documents are to be retrieved in a conventional system by persons who originally perform the indexing task).

---

[*] It has, however, been conjectured that an inverse relationship exists between recall and precision, such that high recall automatically implies low precision and vice versa.[4,5]

In order to calculate the standard recall and precision measures the following important tasks must be undertaken:

(a) relevance judgments must be made by hand in order to decide for each document and for each search request whether the given document is relevant to the given request;

(b) the relevance judgments are usually all-or-nothing decisions, so that a given document is assumed either wholly relevant or wholly irrelevant (in case of doubt relevance is assumed); and

(c) a cutoff in the correlation between documents and search requests is normally chosen, such that documents whose correlation exceeds the cut-off value are retrieved, while the others are not retrieved.

B. The Generation of Relevance Judgments

A great deal has been written concerning the difficulties and the appropriateness of the various operations listed in Part A.[4,5,6,7] The first task, in particular, which may require the performance of hundreds of thousands of human relevance judgments for document collections of reasonable size is extremely difficult to satisfy and to control.

Two solutions have been suggested, each of which would base the relevance decisions on less than the whole document collection. The first one consists in using sampling techniques to isolate a suitable document subset, and in making relevance judgments only for documents included in that subset. However, if the results obtained for the subset are to be applicable to the total collection, it becomes necessary to choose a sample

representative of the whole. For most document collections this turns out to be a difficult task.

The other solution consists in formulating search requests based on specific source documents included in the collection, and in measuring retrieval performance for a given search request as a function of the retrieval of the respective source documents. This procedure suffers from the fact that search requests based on source documents are often claimed to be nontypical, thus introducing a bias into the measurements which does not exist for requests reflecting actual user needs.

Since the document collection used in connection with the present experiments is small enough to permit an exhaustive determination of relevance, the possible pitfalls inherent in the sampling procedure and in the use of source documents were avoided to a great extent. Many of the problems connected with the rendering of relevance judgments are, however, unresolved for general document collections.

C. The Cut-off Problem

The other major problem is caused by the requirement to pick a correlation cut-off value to distinguish retrieved documents from those not retrieved. Such a cutoff introduces a new variable, which seems to be extraneous to the principal task of measuring retrieval performance. Furthermore, in the SMART system, a different cutoff would have to be picked for each of the many processing methods, if it were desired to retrieve approximately the same number of documents in each case.

ROW CORRELATIONS OF PHRASE-DOCUMENT MATRIX SEPTEMBER 28, 1964 PAGE 73

DIFFERNTL EQ 1A COMPUTER 0.1234	DIFFERNTL EQ 384STABILITY 0.6675	0.9800 0
DIFFERNTL EQ 2MICRO-PROGR 0.0875	DIFFERNTL EQ 360SIMULATIN 0.5758	0.9600 0
DIFFERNTL EQ 3THE ROLE OF 0.0293	DIFFERNTL EQ 200SOLUTION 0.5663	0.9400 0
DIFFERNTL EQ 4A NEW CLASS 0.0844	DIFFERNTL EQ 392ON COMPUT 0.5508	0.9200 0
DIFFERNTL EQ 5ANALYSIS OF 0.0658	DIFFERNTL EQ 386ELIMINATI 0.5483	0.9000 0
DIFFERNTL EQ 6GENERALIZED 0.0741	DIFFERNTL EQ 103RUNGE-KUT 0.5444	0.8800 0
DIFFERNTL EQ 7AN IMPROVED 0.2090	DIFFERNTL EQ 85NOTE ON AN 0.4510	0.8600 0
DIFFERNTL EQ 8SHORT-CUT M 0.0861	DIFFERNTL EQ 192SOLVING E 0.4106	0.8400 0
DIFFERNTL EQ 9OPERATION A 0.0611	DIFFERNTL EQ 358STABILIZA 0.3986	0.8200 0
DIFFERNTL EQ 10ACCURATE T 0.1100	DIFFERNTL EQ 102ON THE SO 0.3986	0.8000 0
DIFFERNTL EQ 12DIGITAL CO 0.0883	DIFFERNTL EQ 387BOUNDARY 0.3968	0.7800 0
DIFFERNTL EQ 13HALF-ADDER 0.0548	DIFFERNTL EQ 202STABLE PR 0.3906	0.7600 0
DIFFERNTL EQ 16CONTROL AP 0.0336	DIFFERNTL EQ 229MATRIX PR 0.3505	0.7400 0
DIFFERNTL EQ 17THE FUNCTI 0.0580	DIFFERNTL EQ 88PROPOSED M 0.3451	0.7200 0
DIFFERNTL EQ 18AN ACCURAT 0.1397	DIFFERNTL EQ 251ERROR EST 0.3329	0.7000 0
DIFFERNTL EQ 19RESISTANCE 0.0177	DIFFERNTL EQ 234ANALOGUE 0.3176	0.6800 0
DIFFERNTL EQ 20DIFFERENTI 0.2123	DIFFERNTL EQ 253ROUND-OFF 0.3152	0.6600 1
DIFFERNTL EQ 21AN ERROR-C 0.0105	DIFFERNTL EQ 186ALGORITHM 0.3144	0.6400 1
DIFFERNTL EQ 22LATCHING C 0.0057	DIFFERNTL EQ 169THEORETIC 0.3136	0.6200 1
DIFFERNTL EQ 23MINIATURE 0.0307	DIFFERNTL EQ 128COMPUTER 0.3034	0.6000 1
DIFFERNTL EQ 24SOME NOVEL 0.0199	DIFFERNTL EQ 226+DEPI .. 0.3028	0.5800 1
DIFFERNTL EQ 25A NEW TRAN 0.1068	DIFFERNTL EQ 45A CALCULAT 0.2958	0.5600 3
DIFFERNTL EQ 26SEMICONDUC 0.0653	DIFFERNTL EQ 390MONTE CAR 0.2866	0.5400 6
DIFFERNTL EQ 27TEN MEGAPU 0.1004	DIFFERNTL EQ 388A METHOD 0.2787	0.5200 6
DIFFERNTL EQ 28DESIGN OF 0.1375	DIFFERNTL EQ 173AUTOMATIC 0.2753	0.5000 6
DIFFERNTL EQ 29INVESTIGAT 0.0879	DIFFERNTL EQ 306ELECTRONI 0.2750	0.4800 6
DIFFERNTL EQ 30A TRANSIST 0.0736	DIFFERNTL EQ 318FROM FORM 0.2741	0.4600 6
DIFFERNTL EQ 31MAGNETIC C 0.0575	DIFFERNTL EQ 249MATHEMATI 0.2683	0.4400 7
DIFFERNTL EQ 32ANALOGUE I 0.2283	DIFFERNTL EQ 266UNIFYING 0.2682	0.4200 7
DIFFERNTL EQ 33THE USE OF 0.0802	DIFFERNTL EQ 217SIMULATIO 0.2665	0.4000 8
DIFFERNTL EQ 34END-FIRED 0.0458	DIFFERNTL EQ 367ON EXPONE 0.2661	0.3800 12
DIFFERNTL EQ 35A LOAD-SHA 0.0331	DIFFERNTL EQ 213PREDICTIO 0.2630	0.3600 12
DIFFERNTL EQ 36FUNDAMENTA 0.0392	DIFFERNTL EQ 108SECANT MO 0.2620	0.3400 14
DIFFERNTL EQ 37A HIGH-SPE 0.0364	DIFFERNTL EQ 383A NOTE ON 0.2580	0.3200 15
DIFFERNTL EQ 38AUTOMATIC 0.1043	DIFFERNTL EQ 191DIGITAL C 0.2370	0.3000 21
DIFFERNTL EQ 41COMMUNICAT 0.1185	DIFFERNTL EQ 171SMALL COM 0.2325	0.2800 23
DIFFERNTL EQ 42A DIRECT R 0.0439	DIFFERNTL EQ 248METHOD FO 0.2319	0.2600 33
DIFFERNTL EQ 43THE DATA C 0.0333	DIFFERNTL EQ 283BINARY AR 0.2311	0.2400 34
DIFFERNTL EQ 44ACCURACY C 0.1399	DIFFERNTL EQ 252A CLASS O 0.2303	0.2200 47
DIFFERNTL EQ 45A CALCULAT 0.2958	DIFFERNTL EQ 385NUMERICAL 0.2300	0.2000 60
DIFFERNTL EQ 46RADIO DIRE 0.0980	DIFFERNTL EQ 210EVALUATIO 0.2285	0.1800 73
DIFFERNTL EQ 47SPECIAL PU 0.1268	DIFFERNTL EQ 220DATA PREP 0.2282	0.1600 87
DIFFERNTL EQ 48A BUSINESS 0.0086	DIFFERNTL EQ 32ANALOGUE I 0.2282	0.1400 106
DIFFERNTL EQ 49A DUAL MAS 0.0575	DIFFERNTL EQ 197TECHNICAL 0.2280	0.1200 135
DIFFERNTL EQ 50ACCURACY C 0.0668	DIFFERNTL EQ 355A ROUTINE 0.2272	0.1000 166
DIFFERNTL EQ 52+ATHENA , 0.1030	DIFFERNTL EQ 215DIGITAL C 0.2259	0.0800 200
DIFFERNTL EQ 53A COMPUTER 0.1327	DIFFERNTL EQ 69COMPUTERS 0.2249	0.0600 257
DIFFERNTL EQ 54AN AUTOMAT 0.0763	DIFFERNTL EQ 201ITERATIVE 0.2198	0.0400 304
DIFFERNTL EQ 55AUTOMATIC 0.0746	DIFFERNTL EQ 193ARTIFICIA 0.2196	0.0200 348
DIFFERNTL EQ 56THE COMPUT 0.1513	DIFFERNTL EQ 361SAINT COM 0.2187	
DIFFERNTL EQ 57CASE STUDY 0.0950	DIFFERNTL EQ 257SURVEY OF 0.2161	
DIFFERNTL EQ 58THE LARGES 0.0256	DIFFERNTL EQ 236OPERATING 0.2180	
DIFFERNTL EQ 59DATA PROCE 0.0302	DIFFERNTL EQ 117COMPUTATI 0.2170	
DIFFERNTL EQ 60INTELLIGEN 0.0091	DIFFERNTL EQ 207AN APPLIC 0.2162	
DIFFERNTL EQ 61AN INPUT R 0.0404	DIFFERNTL EQ 20DIFFERENTI 0.2122	
DIFFERNTL EQ 62ON PROGRAM 0.1181	DIFFERNTL EQ 235FREEZING 0.2093	

a) INCREASING DOCUMENT ORDER

b) DECREASING CORRELATION ORDER

c) HISTOGRAM

Correlations Between Search Request and Document Collection

Figure 5

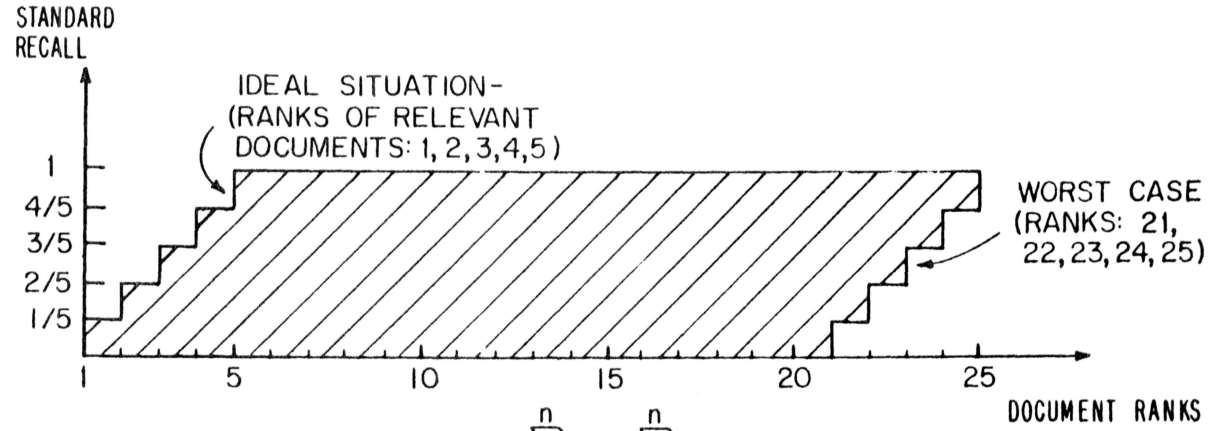
Because of these added complications, it was felt that the standard recall and precision measures should be redefined so as to remove the necessary distinction between retrieved and nonretrieved information. Consider for this purpose the list of documents obtained in answer to a search request, arranged in decreasing order of the correlation coefficients. Such a list is reproduced for the previously used request on differential equations in the center section of Fig. 5. It may be seen that in the figure, document 384 exhibits the longest correlation with the search request, followed by documents 360, 200, 392, and so on. An ordered document list of the kind shown in Fig. 5 suggests that a suitable criterion for recall and precision measures would be the set of rank-orders of the relevant documents, when these documents are arranged in decreasing correlation order. A function of the rank-order list which penalizes high ranks for relevant documents (and therefore low correlation coefficients) can be used to express recall, while a function penalizing low ranks of nonrelevant documents is indicative of precision.

D. Normalized Recall and Normalized Precision[*]

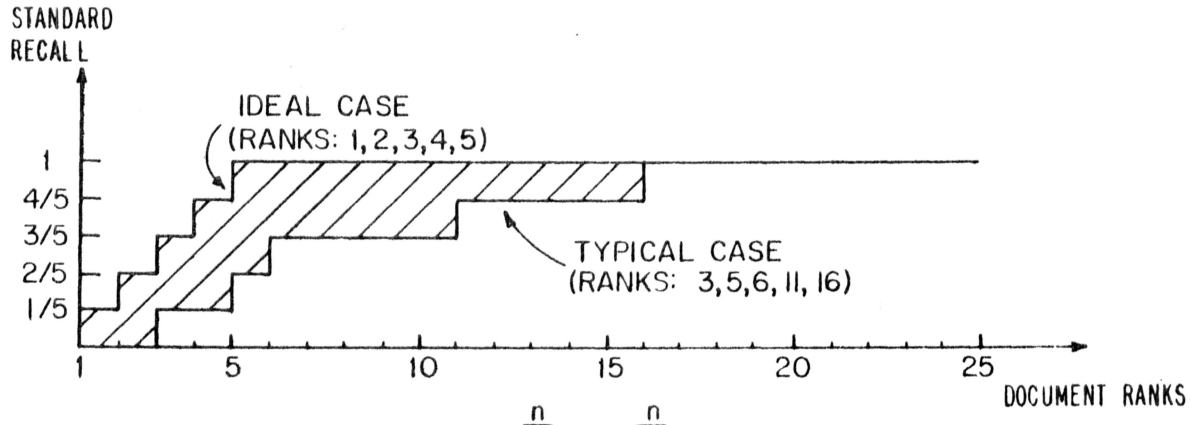
The derivation for the proposed recall measure, called normalized recall, is shown in Fig. 6. The measure is based on the area difference (the integral) between an assumed ideal recall curve, where all relevant documents appear at the top of the ordered list with ranks 1,2,3,..., and

---

[*] The measures described in this section were suggested by J. Rocchio.[8]



$$\text{AREA DIFFERENCE (WORST CASE)} \quad \frac{\sum_{i=1}^{n} r_i - \sum_{i=1}^{n} i}{n} = N - n = 20$$



$$\text{AREA DIFFERENCE (UNNORMALIZED)} \quad \frac{\sum_{i=1}^{n} r_i - \sum_{i=1}^{n} i}{n} = \frac{26}{5} = 5.2$$

$$\text{NORMALIZED RECALL} \quad 1 - \frac{\sum_{i=1}^{n} r_i - \sum_{i=1}^{n} i}{n(N-n)} = 0.74$$

$n = 5$ (NO. OF RELEVANT DOCUMENTS)

$N = 25$ (NO. OF DOCUMENTS IN COLLECTION)

r_i (RANK ORDER OF i^{th} RELEVANT DOCUMENT)

Construction of Normalized Recall Measure

Figure 6

the actual recall curve obtained by plotting the standard recall against the document ranks. If there are n relevant documents, and if r_i is the rank of the ith relevant document, the area difference is clearly

$$\frac{\sum_{i=1}^{n} r_i - \sum_{i=1}^{n} i}{n}.$$

For the case with five relevant documents illustrated in Fig. 6, the ideal ranks of the relevant documents are 1, 2, 3, 4, and 5; the actual assumed ranks shown in the figure are 3, 5, 6, 11, and 16, so that the area difference is 5.2 in that case.

This area difference is, however, not normalized, and its maximum value may increase indefinitely with increasing size N of the document collection. The maximum possible area between the two recall curves is obtained for the worst case, where the relevant documents are ranked N-(n-1), N-(n-2),...,N. In that case the area difference may be seen to be exactly N - n. To generate a normalized measure it is then necessary to divide by N - n, thus obtaining

$$\frac{\sum_{i=1}^{n} r_i - \sum_{i=1}^{n} i}{n(N-n)}.$$

	STANDARD DEFINITIONS (BASED ON THRESHOLD TO DISTINGUISH DOCUMENTS RETRIEVED FROM DOCUMENTS NOT RETRIEVED)	DEFINITIONS BASED ON RANKS OF RELEVANT DOCUMENTS (DOCUMENTS ARRANGED IN ORDER BY DECREASING CORRELATION WITH SEARCH REQUESTS)
RECALL	PROPORTION OF RELEVANT MATERIAL ACTUALLY RETRIEVED (LOW RECALL IMPLIES THAT SOME RELEVANT DOCUMENTS HAVE LOW CORRELATION WITH SEARCH REQUEST AND ARE THUS NOT RETRIEVED)	$1 - \frac{\sum_{i=1}^{n} r_i - \sum_{i=1}^{n} i}{n(N-n)}$ n: NUMBER OF RELEVANT DOCUMENTS N: NUMBER OF DOCUMENTS IN COLLECTION r_i: RANK ORDER OF i^{th} RELEVANT DOCUMENT
PRECISION	PROPORTION OF RETRIEVED MATERIAL ACTUALLY RELEVANT (LOW PRECISION IMPLIES THAT SOME NON-RELEVANT DOCUMENTS HAVE A HIGH CORRELATION WITH SEARCH REQUEST AND ARE THUS RETRIEVED)	$1 - \frac{\sum_{i=1}^{n} \ell n\, r_i - \sum_{i=1}^{n} \ell n\, i}{\ell n \frac{N!}{n!\,(N-n)!}}$

Basic Definitions of Recall and Precision

Figure 7

This measure ranges from 0 for perfect recall to 1 for the worst possible case. A subtraction from 1 now furnishes a measure ranging from 1 to 0 instead of from 0 to 1, with the following expression:

$$R_{norm}\ (\text{normalized recall}) = \quad 1 - \frac{\sum_{i=1}^{n} r_i - \sum_{i=1}^{n} i}{n(N-n)}.$$

A similar derivation for the precision results in the formula:

$$P_{norm}\ (\text{normalised precision}) = \quad 1 - \frac{\sum_{i=1}^{n} \ln r_i - \sum_{i=1}^{n} \ln i}{\ln \frac{N!}{n!(N-n)!}}.$$

These basic definitions are summarized in the table of Fig. 7.

5. Test Results

A. Output Formats

The normalized recall and precision measures are a function only of the ranks of the relevant documents. If these measures are to be evaluated automatically as part of the retrieval process, it is necessary to introduce for each search request processed a list of the corresponding relevant document identifications. To this effect the requestor is given a copy of the full document collection after his request is received, and he is asked to list those documents which he believes should be considered relevant to his request. It is important to note that these relevance

judgments are a priori judgments, based on the document texts only, and not on any retrieval results produced by the computer.

The type of output obtained from the evaluation process is illustrated in Fig. 8. The top part of the figure represents the output from the regular thesaurus procedure for the request on differential equations, while the bottom part is produced by the statistical phrase method. On the right side of the figure appears the list of all 16 relevant document numbers, as originally submitted by the user, together with the respective correlation coefficients and the ranks assigned by the computer during the retrieval process. It may be noticed that the relevant document which exhibits the lowest correlation with the search request is ranked 40th out of 405 by the regular thesaurus procedure, but only 25th out of 405 by the statistical phrase search.

The document ranks are used by the program to produce a variety of measures reflecting recall and precision, including the normalized recall and normalized precision measures previously introduced. Also calculated are simplified expressions, termed respectively rank recall and log precision, and defined as follows:

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

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

EVALUATION OF REQUEST DIFFERNTL EQ WITH 16 RELEVANT DOCUMENTS

THE TOP FIFTEEN DOCUMENTS		RELEVANT DOCUMENT RANKS		
1 X	384STABILITY 0.6676	1	384STABILITY 0.6676	REGULAR THESAURUS
2 X	360SIMULATIN 0.5758	2	360SIMULATIN 0.5758	
3 X	200SOLUTION 0.5664	3	200SOLUTION 0.5664	
4 X	392ON COMPUT 0.5508	4	392ON COMPUT 0.5508	
5 X	386ELIMINATI 0.5484	5	386ELIMINATI 0.5484	
6 X	103RUNGE-KUT 0.5445	6	103RUNGE-KUT 0.5445	
7 X	85NOTE ON AN 0.4511	7	85NOTE ON AN 0.4511	
8	192SOLVING E 0.4106	9	102ON THE SO 0.3987	
9 X	102ON THE SO 0.3987	10	358STABILIZA 0.3986	
10 X	358STABILIZA 0.3986	11	387BOUNDARY 0.3968	
11 X	387BOUNDARY 0.3968	12	202STABLE PR 0.3907	
12 X	202STABLE PR 0.3907	15	251ERROR EST 0.3329	
13	229MATRIX PR 0.3506	17	253ROUND-OFF 0.3152	
14	88PROPOSED M 0.3452	23	390MONTE CAR 0.2866	
15 X	251ERROR EST 0.3329	24	388A METHOD 0.2788	
		40	385NUMERICAL 0.2301	

RANK RECALL= 0.7196 LOG PRECISION= 0.9169
NORMALIZED RECALL=0.9914626 NORMALIZED PRECISION=0.9572
RNK REC + LOG PRE=1.6365 WEIGHTED NORMED RECALL + NORMED PREC=1.9146

THE TOP FIFTEEN DOCUMENTS		RELEVANT DOCUMENT RANKS		
1 X	384STABILITY 0.8576	1	384STABILITY 0.8576	STATISTICAL PHRASE SEARCH
2 X	360SIMULATIN 0.7741	2	360SIMULATIN 0.7741	
3 X	386ELIMINATI 0.7408	3	386ELIMINATI 0.7408	
4 X	392ON COMPUT 0.6571	4	392ON COMPUT 0.6571	
5 X	200SOLUTION 0.6444	5	200SOLUTION 0.6444	
6 X	85NOTE ON AN 0.6372	6	85NOTE ON AN 0.6372	
7 X	387BOUNDARY 0.6072	7	387BOUNDARY 0.6072	
8 X	103RUNGE-KUT 0.5875	8	103RUNGE-KUT 0.5875	
9 X	102ON THE SO 0.5648	9	102ON THE SO 0.5648	
10 X	390MONTE CAR 0.5448	10	390MONTE CAR 0.5448	
11 X	358STABILIZA 0.5437	11	358STABILIZA 0.5437	
12 X	388A METHOD 0.5318	12	388A METHOD 0.5318	
13 X	202STABLE PR 0.5163	13	202STABLE PR 0.5163	
14 X	385NUMERICAL 0.4942	14	385NUMERICAL 0.4942	
15	169THEORETIC 0.4794	21	251ERROR EST 0.3444	
		25	253ROUND-OFF 0.3157	

RANK RECALL= 0.9007 LOG PRECISION= 0.9751
NORMALIZED RECALL=0.9975900 NORMALIZED PRECISION=0.9880
RNK REC + LOG PRE=1.8758 WEIGHTED NORMED RECALL + NORMED PREC=1.9759

Automatic Evaluation

Figure 8

These simple measures are analogous to the normalized recall and normalized precision, but do not take into account the collection size N.

Finally, two composite measures are produced which include both recall and precision components. The first one consists simply of the sum of rank recall plus log precision. The other is a weighted sum of the normalized measures, as follows:

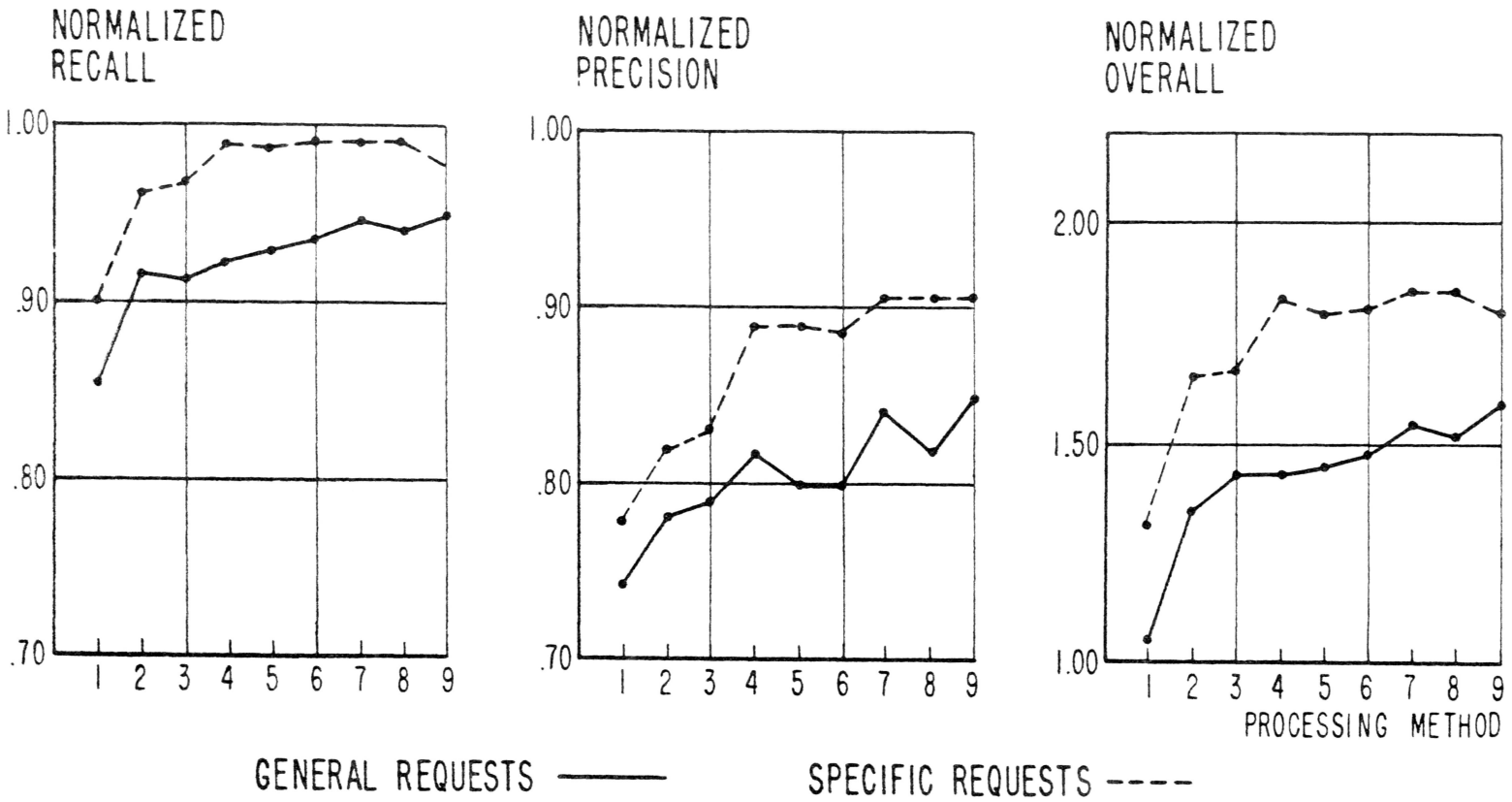
$$\text{normed over-all measure} = 1 - 5(R_{norm}) + P_{norm}.$$

The factor of 5 is so chosen as to give equal weight to the two component measures.

Also included in Fig. 8 are lists of the 15 documents which exhibit the highest correlation coefficients with the search request. The relevant documents on that list are provided with a special marker (X). It may be seen that for the example of Fig. 8, the recall and precision values obtained by a statistical phrase process are larger than the corresponding values for the thesaurus lookup procedure.

B. Results Derived from the Normalized Measures

In order to obtain statistically useful measurements, the recall and precision values must be averaged over many different search requests. This is done in Fig. 9 for nine different processing methods, and for a total of ten specific and seven general requests. A number of obvious conclusions become immediately apparent from the data of Fig. 9:



PROCESSING METHODS:

1. THESAURUS – TITLES ONLY
2. THESAURUS – HIERARCHY (UP AND ADD)
3. THESAURUS – LOGICAL VECTORS
4. WORD STEMS – FULL TEXT
5. THESAURUS – SYNTACTIC PHRASES
6. THESAURUS – HIERARCHY (DOWN AND ADD)
7. THESAURUS – NUMERIC VECTORS
8. THESAURUS – STATISTICAL PHRASES (REQUESTS ONLY)
9. THESAURUS – STATISTICAL PHRASES (WHOLE)

Normalized Recall, Precision and Over-all Measures
(Averaged Over Ten Specific and Seven General Requests
for Several Processing Methods)

Figure 9

(a) the normalized measures obtained for the various processing methods exhibit substantial differences;

(b) as one proceeds from one method to another, both recall and precision tend to vary in the same direction (either up or down);

(c) all the measures (recall, precision, and over-all) obtained for the specific requests are larger than the corresponding values for the general requests, thus indicating a better systems performance for clearly specified logic classes;[†]

(d) methods one to four tend to produce relatively poorer recall than methods five to nine; these same methods also furnish relatively poor precision:

(e) the use of the regular thesaurus which provides vocabulary control (method seven) seems much more effective than the use of the original words included in document and search requests (method four);[‡] and

(f) the most effective procedures seem to be those which use combinations of concepts (phrases), rather than individual concepts alone.

The data of Fig. 9 are of interest in themselves, since they do support the notices that more reasonable procedures (than mere word matching) can be generated to improve retrieval effectiveness in an

---

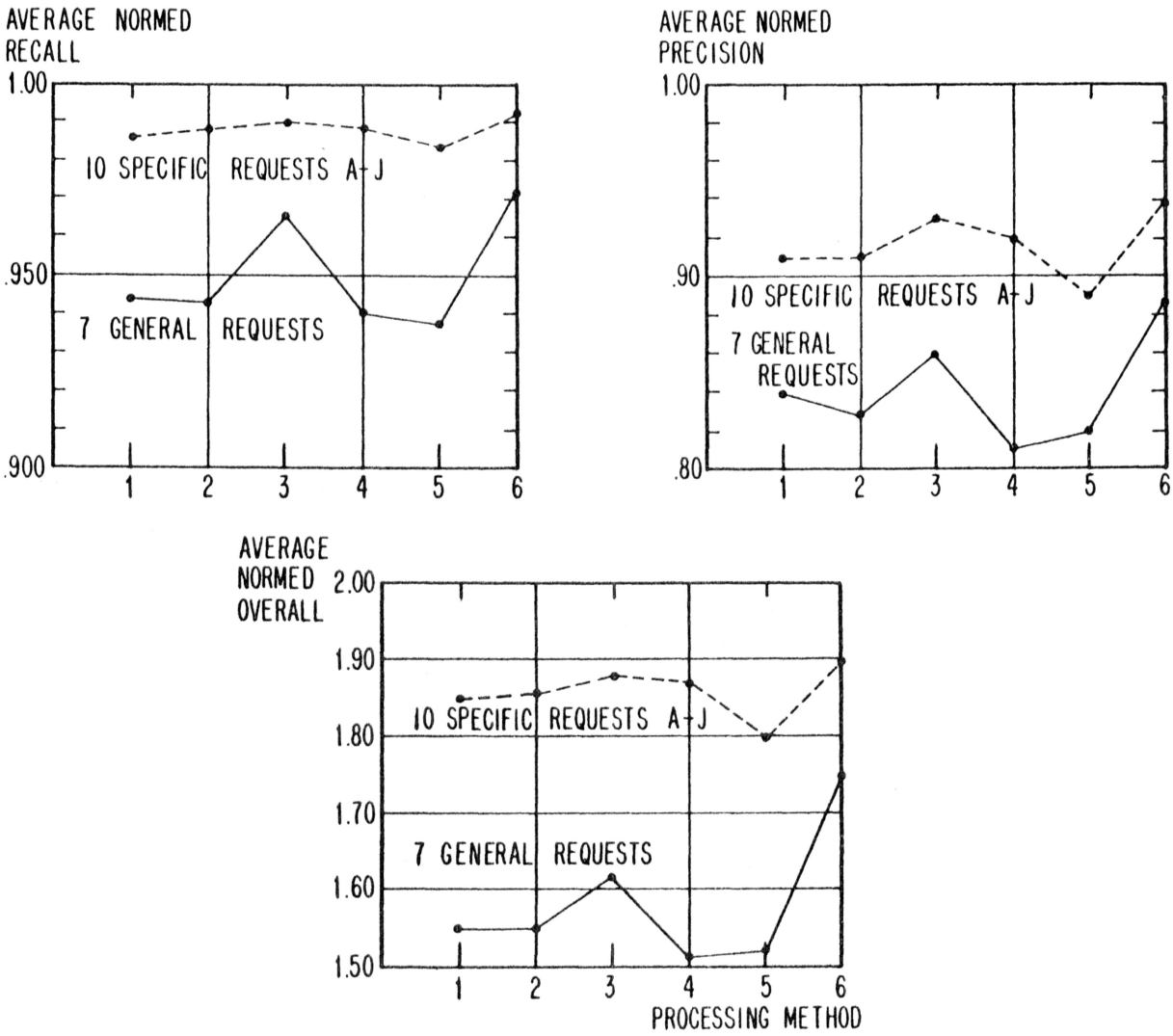
[†] These results would seem to indicate that Cleverdon's observation, reported by Swets,[3] that specific requests will have high precision and low recall, and vice versa for general requests, need not necessarily hold in all circumstances.

[‡] This observation has, of course, been made many times before, particularly by librarians and documentalists, but still requires emphasis in computer circles.

automatic system. However, if full advantage is to be taken of the organization of the SMART system, then search requests are best processed by several different methods, and the respective outputs combined. In order to determine whether this juxtaposition of methods can in fact be used to improve the performance characteristics, average normalized recall and precision figures are given in Fig. 10 for six combined methods and for the requests previously used in Fig. 9.

Figure 10 includes the normalized recall and precision values for the regular thesaurus run previously shown in Fig. 9, followed by the same measures for various combined methods. All of the combined runs include the regular thesaurus run as a component. It may be seen that for three of the combined methods (methods two, three, and six), the over-all measures for both specific and general requests are larger than for any of the included methods alone. Method six, consisting of a combination of regular thesaurus plus word stems plus statistical phrase runs, seems to be particularly effective.

The normalized recall and precision measures for the combined methods are computed by using the rank lists produced by the computer for the individual methods alone, and automatically generating a combined rank list. The combined rank of a given document depends on the individual ranks held by that document in the component methods. Specifically, documents are taken alternately from the component lists to form the new combined list, and a document already included on the combined list is rejected if an attempt is made to list it again. The final combined rank list is then

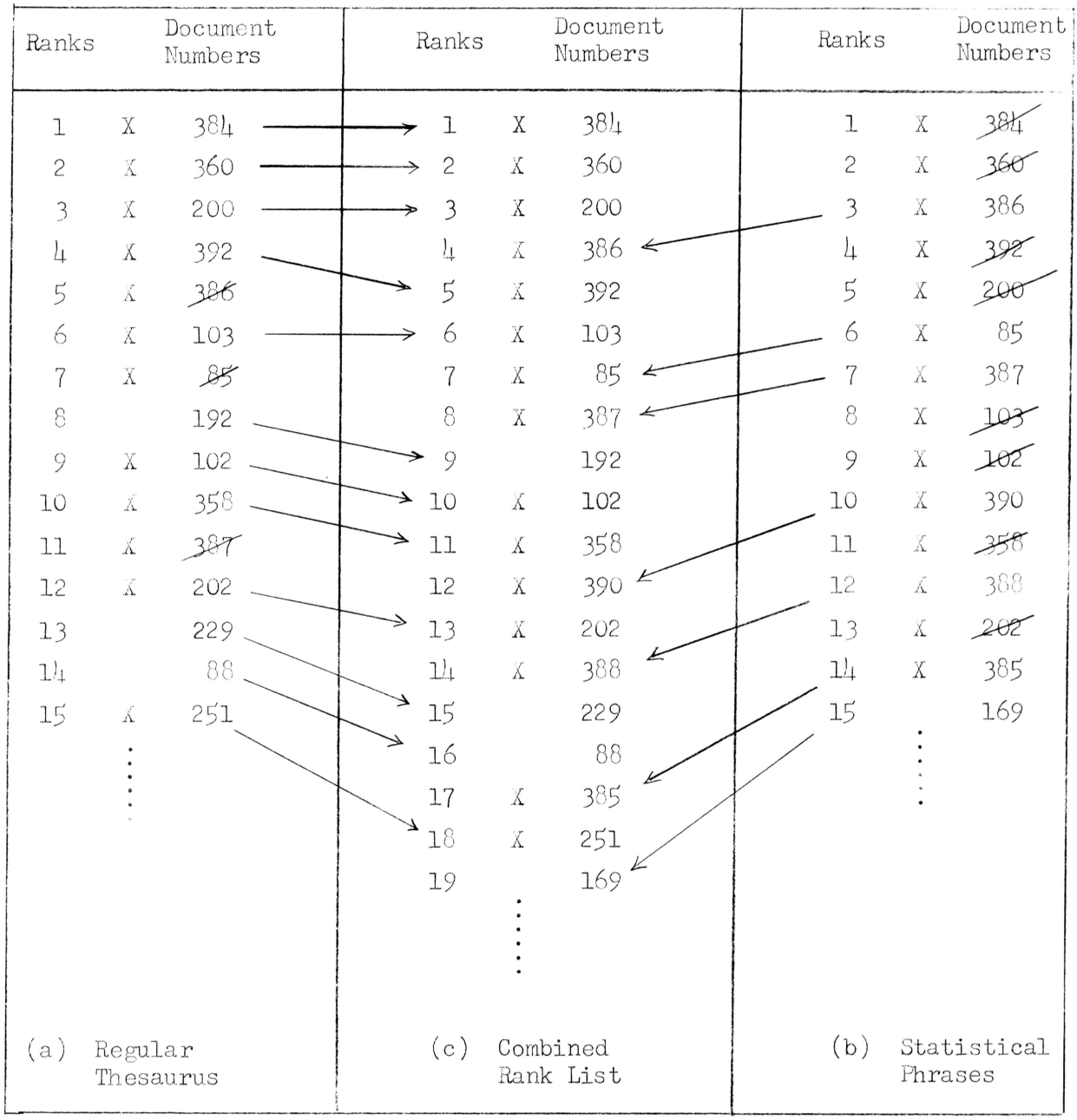


PROCESSING METHODS:

1. REGULAR THESAURUS
2. REGULAR THESAURUS + WORD STEMS
3. REGULAR THESAURUS + STATISTICAL PHRASES
4. REGULAR THESAURUS + HIERARCHY (DOWN)
5. REGULAR THESAURUS + HIERARCHY (UP)
6. REGULAR THESAURUS + WORD STEMS + STATISTICAL PHRASES

Normalized Recall, Precision and Over-all Measures
for Several Merged Methods

Figure 10



Merging of Rank-order Lists

Figure 11

used to compute recall and precision measures for the combined methods, as previously specified in Sec. 4. The resulting measures are averaged over several search requests to produce the graphs of Fig. 10.

A combined rank list, generated for the two methods illustrated in Fig. 8, is shown in Fig. 11 (only the first 15 documents are included for each component method). Documents previously specified as relevant are marked with an X, as in Fig. 8.

C. Results Using the Standard Measures

The performance characteristics of the SMART retrieval operations are reflected with reasonable accuracy in the data of Figs. 9 and 10. In particular, these figures can be used to obtain an idea of the relative effectiveness of one method compared with another. The data are, however, difficult to interpret in absolute terms, particularly since the measures used are new ones, and no comparable output is available elsewhere in the literature.

In order to furnish some indication of systems performance which could lend itself to a comparison with previously published data, the standard recall and precision measures reflecting, respectively, the proportion of relevant material retrieved, and the proportion of retrieved material relevant, are also computed for the search requests previously used. To generate these functions, it becomes necessary to choose appropriate threshold values which separate the retrieved information from that not retrieved. The procedure adopted for this purpose is as follows:

(a) a specified standard recall value is picked (say 0,1);

(b) the number of documents which must be retrieved for a given search request in order to produce the specified recall is determined;

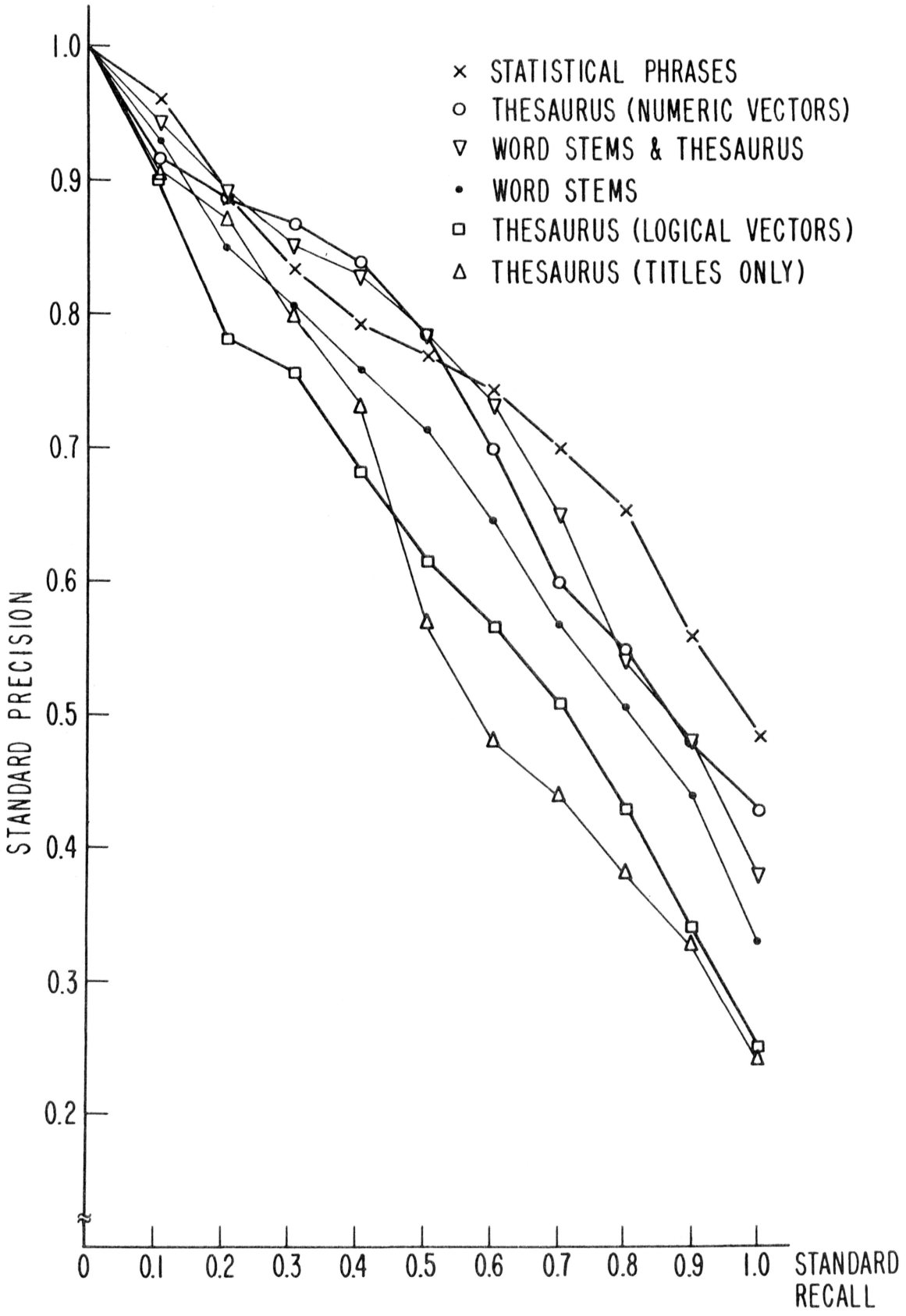
(c) using the value calculated under (b) for the number of retrieved documents, the precision measure (corresponding to the specified recall) is generated;

(d) the precision values obtained for a given recall level are averaged over a number of search requests, and the corresponding point is plotted on a precision versus recall plot; and

(e) the complete procedure is repeated for a new recall level (say 0.2, and 0.3, and so on) to produce a curve of the type shown in Fig. 12.

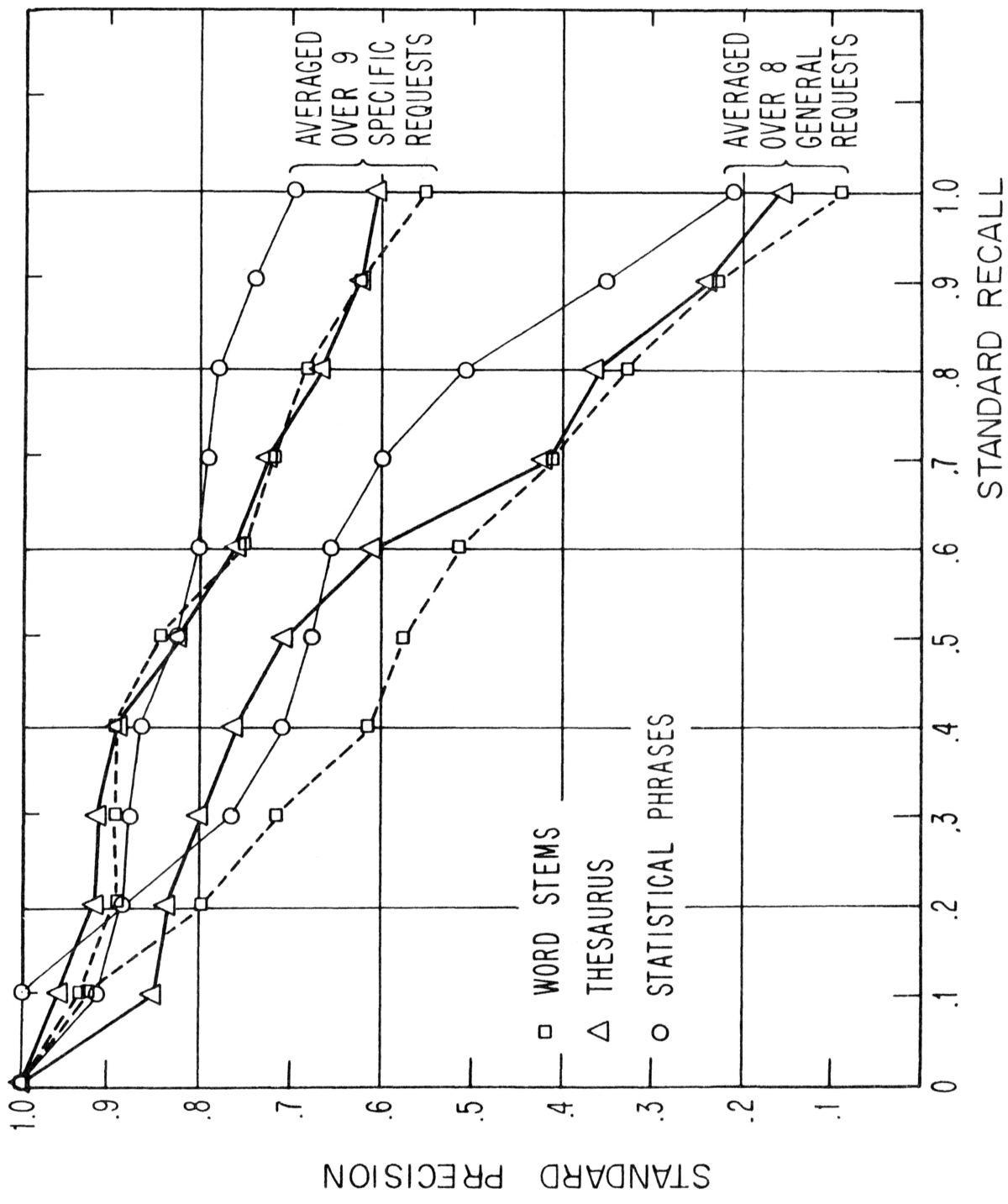
Figure 12 displays the standard precision versus standard recall graphs obtained for six processing methods, averaged over the 17 search requests previously used in Figs. 9 and 10. Figure 12 is in the exact form introduced by Cleverdon,[4,5] using the standard precision and recall measures, rather than the normalized measures based on the rank lists; the procedure previously given to generate the average precision over several requests is believed to be somewhat different from Cleverdon's, but the figures presented should nevertheless lend themselves to a comparison with the published Cranfield material.†

† Recall versus precision plots have been criticized, because important information reflected in separate plots of recall and precision is obscured in the combined presentation (notably the number of documents both retrieved and relevant.[9]



Standard Precision vs. Standard Recall
(Average Values Over 17 Requests)

Figure 12



Standard Precision vs. Standard Recall
Comparison of General and Specific Requests

Figure 13

The data of Fig. 12 confirms those previously shown in Fig. 9 in that the statistical phrase run again seems to give the best performance. Furthermore, word stem comparisons are again inferior to the regular thesaurus runs, and "titles only" analysis is generally inferior. The differences in systems performance previously noted for the output of Figs. 9 and 10 are again in evidence, since for a given recall level, average precision can vary by over 35 percent from one method to another. The same is true of the average recall differences for a given level of precision.

Figure 13 shows standard precision versus standard recall figures averaged separately over the specific and the general requests for three processing methods. A comparison with Fig. 10 again indicates that both recall and precision measures are substantially higher for the specific requests than for the general requests.

6. Conclusions

The evaluation procedures and results included in the present study are based on the manipulation of one relatively small collection of document abstracts, and a set of about 20 search requests. Only about 15 different processing methods are used. Under the circumstances, it is not possible to make claims of general validity or to prove many assertions with finality.

Nevertheless, it is believed that the data presented here can be used as indications of the kind of performance to be expected of automatic retrieval systems. In particular, the data which point to the existence of considerable discrepancies in performance characteristics between

processing methods may be expected to be confirmed by new experiments with different document collections and larger numbers of search requests. Of special interest in this connection is the fact that certain processing methods exhibit both high recall and high precision, thus indicating good over-all performance.

The other principal piece of evidence tends to support the notion that the juxtaposition of a variety of processing methods provides improved retrieval performance over and above the performance of the individual component methods. The design philosophy of the SMART system, which is based on an iterative search procedure with a variety of analysis methods to retrieve relevant information, should therefore prove useful in practice. (A similar conclusion, pointing to the joint use of UDC (universal decimal classification) coupled to a Uniterm system, has previously been reached in a conventional retrieval situation.)[10]

Additional experiments remain to be carried out with different document collections not previously used with the available dictionaries, and with additional search requests. A careful analysis of systems failures is also mandatory, in order to determine more precisely the strengths and weaknesses of the individual methods, and the circumstances under which relevant documents are not recognized, and receive therefore a low correlation on the output lists. Additional processing sequences must also be analyzed and useful sequences identified, in order to maximize system performance and retrieval effectiveness.

REFERENCES

1. Salton, G., "A Document Retrieval System for Man-Machine Interaction," Proceedings of the 19th ACM Annual Conference, Philadelphia (1964).

2. Salton, G., et al., Information Storage and Retrieval, Report No. ISR-7 to the National Science Foundation, Computation Laboratory of Harvard University (June 1964).

3. Swets, J. A., Information Retrieval Systems, Science, Vol. 141, No. 3577 (July 19, 1963).

4. Cleverdon, C. W., "The Testing of Index Language Devices," ASLIB Proceedings, Vol. 15, No. 4 (April 1963).

5. Cleverdon, C. W., "The Testing and Evaluation of the Operating Efficiency of the Intellectual Stages of Information Retrieval Systems," International Study Conference on Classification Research, Elsinore (September 1964).

6. Swanson, D. R., "Searching Natural Text by Computer," Science, Vol. 132, No. 3434 (October 21, 1960).

7. Swanson, D. R., "Design of Experiments for the Testing of Indexing Systems," draft manuscript.

8. Rocchio, J., "Performance Indices for Document Retrieval Systems," Information Storage and Retrieval, Report No. ISR-8 to the National Science Foundation, Computation Laboratory of Harvard University (December 1964).

9. Fairthorne, R. A., "Basic Parameters of Retrieval Tests," short paper presented at the 1964 ADI Annual Meeting, Philadelphia (October 1964).

10. Schuller, J. A., "Experience with Indexing and Retrieving by UDC and Uniterms," ASLIB Proceedings, Vol. 12, No. 11 (November 1960).