The Evaluation of Automatic Retrieval Procedures— Selected Test Results Using the SMART System*

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 report 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, and the measures used to assess the effectiveness of the retrieval performance are described. The main test results are given and tentative conclusions are reached concerning the design of fully automatic information systems.

GERARD SALTON

The Computation Laboratory of Harvard University
Cambridge, Massachusetts

• 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 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 lay-out, 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.

• The SMART Retrieval System

SMART is a fully automatic document retrieval system operating on the IBM 7094. Unlike other computer-based retrieval systems, the SMART system does

* This study was supported by the National Science Foundation under Grant GN-245.

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 texts of documents. Instead, an attempt is made to go beyond simple word-matching procedures by using a variety of intellectual aids in the form of synonym dictionaries, hierarchical arrangements of subject identifiers, statistical and syntactic phrase-generating methods and the like, in order to obtain the content identifications useful for the retrieval process.

Stored documents and search requests are then processed *without any prior manual analysis* by one of several hundred automatic content analysis methods, and those documents which most nearly match a given search request are extracted from the document file in answer to the request. The system may be controlled by the user in that a search request can be processed first in a standard mode; the user can then analyze the output obtained and, depending on his further requirements, order a reprocessing of the request under new conditions. The new output can again be examined and the process iterated until the right kind and amount of information are retrieved.

SMART is thus designed to correct many of the shortcomings of presently available automatic retrieval systems, and it may serve as a reasonable prototype for fully automatic document retrieval. The following facilities incorporated into the SMART system for purposes of document analysis may be of principal interest*:

(a) a system for separating English words into *stems* and *affixes* (the so-called "null thesaurus" method) which can be used to construct document identifications consisting of the word stems contained in the documents;

(b) a synonym dictionary, or *thesaurus*, which can be used to recognize synonyms by replacing each word stem by one or more "concept" numbers (the thesaurus is a manually constructed dictionary including about 600 concepts in the computer literature, corresponding to about 3000 English word stems); these concept numbers can serve as content identifiers instead of the original 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; the hierarchy can be used to obtain more general content identifiers than the ones originally given by going "up" in the hierarchy, more specific ones by going "down" in the structure, and a set of related ones by picking up brothers and 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; the related concepts, determined by statistical association, can then be added to the originally available concepts to identify the various documents;

(e) *syntactic analysis* and matching methods which make it possible to compare the syntactically analyzed sentences of documents and search requests with a pre-coded dictionary of "criterion" phrases in such a way that the same concept number is assigned to a large number of semantically equivalent, but syntactically quite different constructions (e.g. "information retrieval," "the retrieval of information," "the retrieval of documents," "text processing," and so on);

(f) *statistical phrase* matching methods which operate like the preceding syntactic phrase procedures, that is, by using a preconstructed dictionary to identify phrases used as content identifiers; however, no syntactic analysis is performed in this case, and phrases are defined as equivalent if the concept numbers of all components match, regardless of the syntactic relationships between components;

(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 type of correlation function 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 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.

*More detailed descriptions of the systems organization are included in Refs. 1 and 2. Programming aspects and complete flowcharts are presented in Ref. 3.

Characteristic	Comment	Count
Number of documents in collection.	Document abstracts in the computer field.	405
Number of search requests (a) specific (b) general.	 0 – 9 relevant documents 10 – 30 relevant documents	 10 7
User population (requester 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 (b) general.		 (average) 5 (average) 15
Number of retrieved documents per request.	No cut-off is used to separate retrieved from nonretrieved.	405

FIG. 1. Test Environment.

• 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 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 fulfilment.

An excerpt from the document collection, as it is originally introduced into 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 is 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 12-character identifier is shown for each document under the heading "answer," and full bibliographic citations are shown under "identification."

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 word stem analysis (labelled "null thesaurus" in Fig. 4); these 35 stems, however, give rise to 50 different concept numbers using the regular thesaurus, and to 55 concepts for the statistical phrase method. The number of index terms per document shown in the summary of Fig. 1 (35) must therefore be taken as an indication at best, and does not properly reflect the true situation.

In Fig. 4, each concept number is followed by some mnemonic characters to identify the concept and by a

* 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.

*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 .

FIG. 2. Typical Document Prints.

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)

FIG. 3. Typical Search Request and Corresponding Answers.

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

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	9LOCAT 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	332SEF 12	338MCH 8	340LET 3	
	346JET 6	350IFO 6	419GEM 6	501URD 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	8ALGUR 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	STAT. PHRASES
	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	332SEF 48	338MCH 8	340LET 3	
	346JET 6	350IFO 6	419GEM 6	501ORD 4	508ACT 6	

PAGE 17

FIG. 4. Typical Indexing Products for Three Analysis Procedures.

weight. The weights assigned to the concept numbers also change from method to method. 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.

• Evaluation Measures

1. 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.[4] Perhaps the best known of these are, respectively, *recall* and *precision*; *recall* is 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 missed at the same time, thus depressing recall. Ideally, one would of course hope for both 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;

* Precision has also been called "relevance," notably in the literature of the ASLIB–Cranfield project.[5]

† 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.

DOCKET ALARM

Explore Litigation Insights

Docket Alarm provides insights to develop a more informed litigation strategy and the peace of mind of knowing you're on top of things.

Real-Time Litigation Alerts

Keep your litigation team up-to-date with **real-time alerts** and advanced team management tools built for the enterprise, all while greatly reducing PACER spend.

Our comprehensive service means we can handle Federal, State, and Administrative courts across the country.

Advanced Docket Research

With over 230 million records, Docket Alarm's cloud-native docket research platform finds what other services can't. Coverage includes Federal, State, plus PTAB, TTAB, ITC and NLRB decisions, all in one place.

Identify arguments that have been successful in the past with full text, pinpoint searching. Link to case law cited within any court document via Fastcase.

Analytics At Your Fingertips

Learn what happened the last time a particular judge, opposing counsel or company faced cases similar to yours.

Advanced out-of-the-box PTAB and TTAB analytics are always at your fingertips.

API

Docket Alarm offers a powerful API (application programming interface) to developers that want to integrate case filings into their apps.

LAW FIRMS

Build custom dashboards for your attorneys and clients with live data direct from the court.

Automate many repetitive legal tasks like conflict checks, document management, and marketing.

FINANCIAL INSTITUTIONS

Litigation and bankruptcy checks for companies and debtors.

E-DISCOVERY AND LEGAL VENDORS

Sync your system to PACER to automate legal marketing.