AUTOMATIC INFORMATION ORGANIZATION AND RETRIEVAL

McGRAW-HILL COMPUTER SCIENCE SERIES

RICHARD W. HAMMING
Bell Telephone Laboratories

EDWARD A. FEIGENBAUM
Stanford University

HELLERMAN Digital Computer System Principles
LIU Introduction to Combinatorial Mathematics
ROSEN Programming Systems and Languages
SALTON Automatic Information Organization and Retrieval
WEGNER Programming Languages, Information Structures, and Machine Organization

Automatic Information Organization and Retrieval

Gerard Salton
Professor of Computer Science
Cornell University

McGraw-Hill Book Company
New York St. Louis San Francisco Toronto London Sydney

Automatic Information Organization and Retrieval

Library of Congress Catalog Card Number 68-25664

07-054485-9

7891011KPKP7987654

Preface

Information retrieval is a field concerned with the structure, analysis, organizatiou, storage, searching, and retrieval of information. In recent years, this whole subject has received an increasing amount of attention from a growing proportion of the technically oriented population, not only because a simplification of the information handling problems becomes ever more urgent, but also because the use of modern computing equipment and sophisticated language processing methods appears to provide the necessary means for generating acceptable solutions.

This book deals with the computer processing of large information files, with special emphasis on automatic text handling methods. Described in particular are procedures for dictionary construction and dictionary look-up, statistical and syntactic language analysis methods, information search and matching procedures, automatic information dissemination systems, and methods for user interaction with the mechanized system. As such, the text includes elements of linguistics, mathematics, and computer programming.

Although none of the chapters requires more than an elementary knowledge of these background subjects, the book is addressed principally to readers who may already have some knowledge of computer processing. It constitutes both a monograph for the professional practitioner versed in general computer utilization and a textbook for the mathematically more mature students who may be enrolled in applied mathematics or computer science curricula or in one of the newly oriented schools of library science.

No attempt is made to survey the entire field of library or information science; indeed, the text deemphasizes some of the more standard subjects such as the description of specialized hardware for information handling. However, the principal developments in mechanized information processing are examined, and the methods given reflect the present state of the art and indicate future trends. For the most part, the material is treated without reference to existing operational systems, and the theories and procedures that are outlined may be applicable to many different computer-based text processing systems. In certain chapters—notably 2, 3, 5, and 8—the experimental Smart document retrieval system serves for illustrative purposes. The Smart system is used because it is the only presently available, computer-based document retrieval system that incorporates sophisticated, fully automatic information analysis procedures in addition to the usual search and retrieval capability. As such, the system may serve as an example of the developments likely to take place in the foreseeable future and as a convenient vehicle for introducing automatic text processing systems.

Chapter 1 is an introduction to information processing and also contains a summary of the Smart system operations. Chapter 2 covers manual and automatic procedures for the construction of a variety of dictionaries and thesauruses useful for the analysis and normalization of natural language texts; covered are word stem dictionaries, thesauruses providing synonym recognition, and various types of phrase dictionaries. Dictionary operations including setup, search, and updating are covered in Chapter 3. Statistical text processing methods are examined in Chapter 4, with emphasis on procedures for the generation of statistical term associations and for the construction of automatic term and document classification systems. Syntactic language analysis operations are covered in Chapter 5, including a variety of procedures for performing automatic syntactic analyses of incoming texts, as well as phrase matching methods to be incorporated into automatic content analysis systems.

Chapter 6 deals with abstract models of the retrieval process and summarizes various attempts at deriving theories of information retrieval; the main emphasis is placed on algebraic models based on set

theory. Practical retrieval operations are described in Chapter 7, including procedures based on partial searches of the stored collections as well as user-controlled iterative search techniques utilizing feedback information supplied by the users of the system. The evaluation of computer-based retrieval systems is covered in Chapter 8; the evaluation results presented are based on the manipulation of document collections in three subject areas: computer science, aerodynamics, and documentation. An attempt is also made to forecast the design of future automatic information systems. Several different classes of auxiliary information services are described in Chapter 9, including systems for the production of index and abstract volumes and methods for the selective dissemination of information. Finally, Chapter 10 is devoted to an examination of retrieval systems based on the storage of structured data bases in restricted subject fields. Various kinds of file manipulations of interest in management information systems are described, and the problems inherent in the design of automatic question answering systems are examined. Future, conversational on-line information systems are also described. Two appendixes cover the detailed operating procedures used with the Smart system and give a selective bibliography.

The various chapters can be examined independently of each other, and a selection best suited to the background and interests of individual readers can be made. A pertinent grouping for the mathematically oriented reader would include Chapters 4 to 8 and 10; for the reader interested in linguistics and natural language processing, Chapters 2, 3, 5, 7, 9, and 10 are of most interest. Either of these sequences could be used for an introductory one-semester course in information organization and retrieval; the former selection has, in fact, been used as part of the graduate curriculum in applied mathematics and computer science at Harvard and Cornell Universities.

As usual, many people deserve credit for the work described in this volume. The author has profited from discussions with many colleagues and friends at Harvard, Cornell, and elsewhere. The work of Professor Susumu Kuno of Harvard has greatly influenced the discussion on automatic syntactic analysis methods, and Cyril W. Cleverdon, Librarian of the Cranfield College of Aeronautics in England, has shaped the material on system evaluation through many days of joint work and much fruitful interaction over several years.

The Smart system would never have been implemented without the exceptional competence and devoted work of Michael E. Lesk, a graduate student at Harvard, who programmed much of the original system and kept the system going on the Harvard 7094 computer after the author moved to Cornell in 1965. Fundamental contributions to the system

were also made by Dr. E. H. Sussenguth, Jr., now at IBM Corporation; Dr. J. J. Rocchio, Jr., now at Bellcomm Incorporated; and E. M. Keen, now at the College of Librarianship in Aberystwyth, Wales.

Several readers who saw the text prior to publication made a number of useful suggestions, including Dr. Don R. Swanson, Dean of the Graduate Library School at the University of Chicago; Mrs. D. Kathryn Weintraub, also of the Graduate Library School at Chicago; Professor Robert M. Hayes of the Institute of Library Research at the University of California in Los Angeles; Herbert R. Koller of Leasco Systems and Research Corporation; and Professor F. P. Brooks, Jr., of the Department of Information Science at the University of North Carolina in Chapel Hill.

Finally, the manuscript could not have been prepared without the assistance of Mrs. Margaret Dodd and Mrs. Sally Grove who typed several drafts cheerfully and competently.

To all these individuals the author is deeply indebted for encouragement, guidance, and help.

Gerard Salton

Contents

1
Automatic Information Systems

1-1 INTRODUCTION

This book deals with the problems of scientific information in the modern world, including its generation and collection, its structure and organization, its analysis, its storage and retrieval, and its dissemination. That there are substantial problems in the information field, everyone is agreed upon: More and more information is generated and put into circulation; the existing tools, classification schedules, and storage arrangements are often inadequate, particularly in the newer fields; and it generally becomes more difficult and more expensive to get to know what one needs to know.

This situation is reflected in the variety of pressures on the administrations of information handling organizations: Budgeting problems become more severe every year; staff positions are increasingly more difficult to fill, particularly in areas requiring specialized skills and know-how; and many of the intellectual problems appear intractable in the present environment. Some experts predict that it may eventually become

so onerous to locate a wanted item of information as to discourage even an attempt to initiate a search:

> As each visit to the library adds to the accumulated annoyances of the user without producing an acceptably high yield of information, the crisis begins to display an ominous symptom: . . . users are staying away from the library [1].[1]

This general climate has helped to foster the notion that new techniques and modern computing equipment may be capable of alleviating and solving to some extent the so-called "information problem." Specifically, many people now believe that there exists the required capacity to store many data or document collections of interest, that procedures are available for analyzing and organizing the information in storage, and that real-time software and hardware can be used to ensure 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.

Dean Shera's opinion is only one of many from within and without the library field:

> . . . there appears to be very good reason for . . . the assumption that machines can be built which will relieve the scholar of much of the burden of bibliographic search, and that they will eventually be able to provide the precise information the user needs when he needs it . . . [2].

Such pronouncements must be counterbalanced by pointing out that the present reality is not up to the promise, in that "the difficult analytical problems inherent in automatic information handling have so far limited the use of automatic techniques in information storage and retrieval to applications which never required much analytical judgment on the part of the humans who formerly did the work [3]."

Be that as it may, a variety of plans have been advanced for the establishment of fully or partly mechanized information and library centers, and recommendations have been drawn up for the organization of national or international document handling systems [4, 5, 6].

This text, then, is an attempt to examine the principal technical and intellectual problems arising in information processing and to determine the extent to which they are amenable to solution by automatic or semiautomatic methods. The structure and properties of scientific information are of principal concern, as reflected in a semantic content analysis

[1] Numbers in brackets indicate numbered references at the end of each chapter.

of the documents (but not a qualitative evaluation concerning their accuracy, veracity, or conciseness) [7].

In many ways, a study of the analytical aspects of scientific information in a mechanized environment must appear as a hopeless endeavor, because so many of the important theoretical problems are unsolved. What exactly is the content or meaning of a document? What are the linguistic devices used to carry meaning? To what extent can individual words, or word groups, in a text be said to carry and maintain a well-defined, controlled meaning? How can one isolate the content-bearing units if they exist? And so on.

Since the answers to these fundamental questions are unclear, it becomes impossible to justify the text-manipulating procedures introduced in this volume for purposes of content analysis other than as ad hoc devices. In other words, although it is reasonable to describe methods such as statistical word associations or syntactic phrase generation and to conduct tests to find out how well these methods operate in a retrieval environment, it is not possible to conclude, even for those procedures that appear to give a satisfactory performance, that they are essential in a content analysis system or to define their exact role as part of such a system.

In introducing many of the topics of interest in automatic information handling, it is necessary to ignore some of the theoretical problems. At the same time, many up-to-date automatic techniques for analyzing, storing, correlating, searching, and retrieving information can be examined, and it can be shown that these techniques are fully equivalent in effectiveness to established manual methods. The present practice, which consists in restricting the mechanization to the file search only and in performing most of the intellectual tasks of document and request analysis on a manual basis, may then appear as an interim step that is not likely to be maintained in the information systems of the future.

This text is concerned with the *structure of information* and with methods for storing word strings in such a way as to make them accessible most efficiently; with *file search* techniques and retrieval programs; with *content analysis* methods based on stored dictionaries as well as on the statistical properties of written texts and on syntactic analysis procedures; with *evaluation* systems designed to monitor the effectiveness of many of the proposed procedures; with *auxiliary outputs* of many scientific information activities, including indexes, abstracts, citation processing, and selective dissemination; and with *question answering* systems based on restricted data bases of the type now considered for many management information applications.

The treatment is based on elements of statistics, linguistics, algebra, and computer programming, but these topics are covered only to the

extent necessary to make the remaining material comprehensible. If and when a theory covering scientific information activity is developed, these elements are likely to be included, in addition to others yet to be developed.

In the remainder of this chapter, an information universe is briefly examined, starting with the generation of information and ending with its utilization by a variety of information handling systems. Finally, an overview is given of the experimental, automatic Smart information storage and retrieval system, since this system is later used as an aid in introducing some of the topics to be covered.

1-2 INFORMATION DISSEMINATION

An information dissemination system may be used to assist in and to control, to some extent, the generation, recording, analysis, classification, storage, search, and retrieval of information. Information in this context may consist of data items, such as *facts* or measurements, or it may consist of written texts, *documents*, books, summaries, abstracts, titles, and so on. The information dissemination process is best described in terms of three main components: information generation, information processing, and information utilization. A stylized representation of the process is shown in Fig. 1-1. The first component, information generation, consists of the steps that are necessary to put a manuscript into final form for publication, including, in particular, all negotiations among authors, editors, reviewers, publishers, and so on. This aspect of the information flow is presently not a part of any formalized system but is

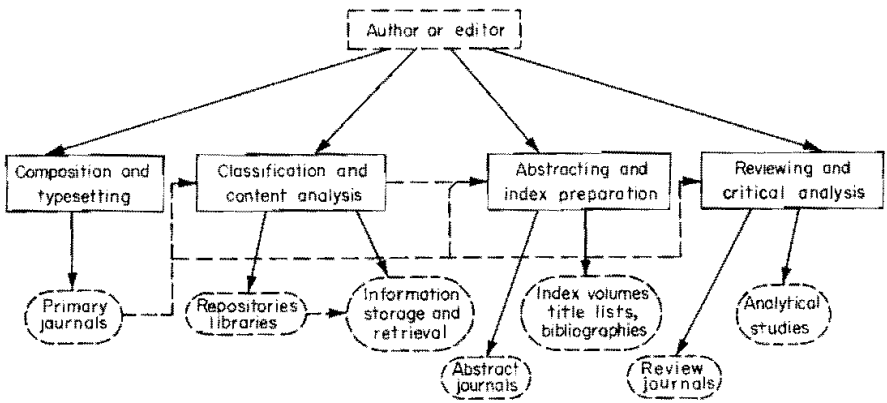


Fig. 1-1 Information generation and dissemination.

handled by personal communication between the parties involved. In the chart of Fig. 1-1, the generation component is represented by the single box at the top of the figure.

The processing component on the second level is represented by four main parts: composition and typesetting tasks necessary to issuing a manuscript in journal form; classification and content analysis prior to introducing a document into an information store; abstracting and indexing to permit inclusion in a variety of secondary publications; and, finally, reviewing and analysis tasks required for review purposes and for other special reasons.

The user interaction component suggested on the third level of Fig. 1-1 comprises a variety of complicated interactions between a user population, a set of information dissemination media such as journals, bulletins, announcements, etc., and a number of information centers established for the purpose of aiding the communication process between originators and users of information.

At the present time, the various tasks that make up the information universe are performed to a large degree independently of each other—often without regard to the total picture. Thus, a manuscript is typed on a keyboard device by the author, and again during the typesetting operation, and again for indexing or abstracting purposes, and again in some instances prior to inclusion in a document storage center. Eventually, one may expect that these related activities may be coordinated in such a way that duplications and overlapping activities are eliminated, for example, by basing all input operations on a unique product derived from a single keyboarding operation.

The information dissemination process is also complicated by the fact that at the present time a variety of organizations take part in the various information activities whose function is often regarded as quite distinct. One can recognize at least three types of information centers as summarized in Table 1-1 [8]:

1. Document depots and libraries of various kinds whose function is to acquire and store selected materials and to make them available in several different forms to a user population
2. Abstracting and indexing services, normally restricted in scope to certain subject areas, whose function includes specifically the preparation and dissemination of abstracts and indexes
3. Information analysis centers charged with the study in depth of certain subject areas and with the preparation of analytical studies

Present plans for the integration of these several activities into a unified system are somewhat indefinite. Because of the existing com-

Table 1-1 Types of information centers

Information center	*Function*
Depots and libraries	Acquisition Storage Reference searching Retrieval Hard copy or microcopy dissemination Abstract dissemination Preparation of bibliographies
Abstracting and indexing service	Acquisition Storage (selective) Abstract preparation and dissemination Index preparation and dissemination
Information analysis center	Acquisition Storage (selective) Reference searching Retrieval Answering of technical questions Preparation and dissemination of analytical studies

petences in many organizations, it is often suggested that the established organizations should maintain their identity but should eventually be integrated into a *network* of information centers such that, from each point, access may be obtained to all available information [9, 10, 11]. Various kinds of system blueprints have been discussed, and some of the organizational questions have generated considerable heat—notably the problem of financing and the related question of government vs. private control, as well as the problems of geographical distribution of centers.

If a unified system of information centers is, in fact, to be implemented, it is likely that some type of discipline oriented plan will initially be followed, first because several document handling systems already exist for special subject areas with experience in their own field, and second because pilot operations are more easily initiated if the coverage is limited and the procedures properly circumscribed. In addition, systems of regional cooperation will undoubtedly be initiated where organizations with large resources will offer services to smaller centers within their region.

1-3 INFORMATION SEARCH AND RETRIEVAL

Because of their special importance in the present context, it is useful to describe in more detail the operations that lead to the retrieval of stored information in answer to user search requests. In practice, searches often may be conducted by using author names or citations or titles as principal criteria. Such searches do not require a detailed content analysis of each item and are relatively easy to perform, provided that there is a unified system for generating and storing the bibliographic citations pertinent to each item.

When the search criteria are based in one way or another on the contents of a document, it becomes necessary to use some system of content identification, such as an existing subject classification or a set of content identifiers attached to each item, which may help in restricting the search to items within a certain subject area and in distinguishing items likely to be pertinent from others to be rejected. In most of the semimechanized centers where the search operation is conducted automatically, it is customary to assign to documents and search requests alike a set of content identifiers, normally chosen from a controlled list of allowable terms, and to compare the respective lists of content identifiers in order to determine the similarity between stored items and requests for information. A simplified chart of the search and retrieval operations is shown in Fig. 1-2.

After assignment of content indicators—an operation usually conducted by human subject experts—the assigned terms are normally validated by comparison with an existing authority list. The acceptable terms assigned to the documents are then compared with the request terms, and logical combinations of terms are generated for the system, capable of retrieving, for example, items dealing with topic A but not with B, or items dealing with either A or B or with C or D. If several searches are run in parallel, a sorting operation is required to separate the items retrieved in response to each request, and the actual bibliographic information corresponding to each retrieved item must be withdrawn from the file and presented to the users.

A large variety of search strategies may be used to obtain satisfactory answers to search requests. Where the user group is heterogeneous, a single search performed for each customer may not suffice to retrieve from the file the exact information that is wanted. Under these circumstances, a sequence of search operations might be carried out in an attempt to approach the desired subject area little by little. This could be done by using information obtained as a result of an earlier search to improve subsequent operation.

A feedback loop is used on both sides of the expanded chart of

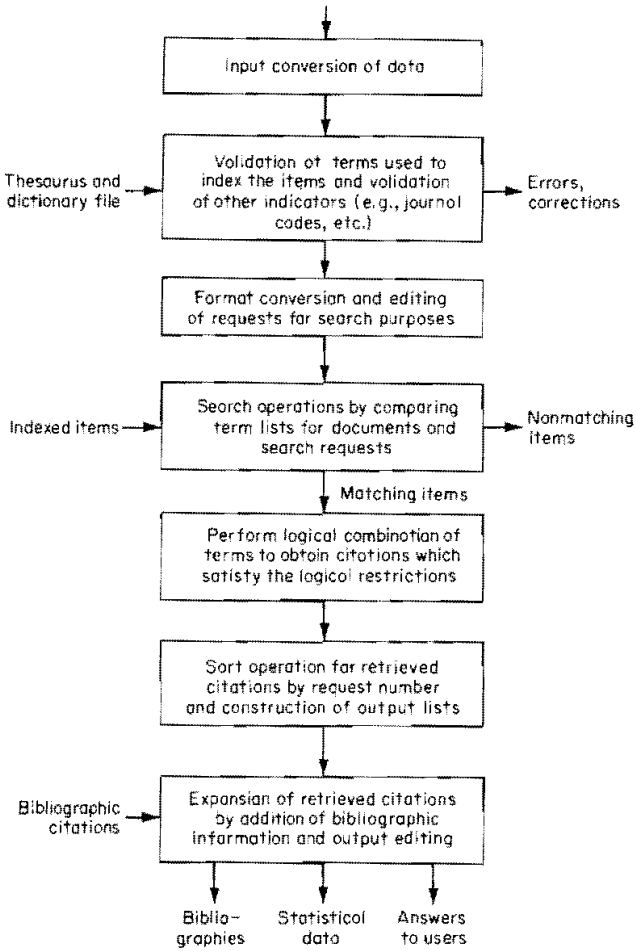


Fig. 1-2 Simplified search and retrieval process.

Fig. 1-3 to represent an iterative search system where the original search requests are updated by using information supplied to the system following an initial search. The left-hand side of Fig. 1-3 represents an ordinary *on-demand* search system where searches are made only after receipt of a request from one or more users; the right-hand side represents *recurrent* searches performed, for example, in a selective dissemination system, where user interest profiles, permanently stored for a group of participating users, take the place of ordinary search requests and are thus compared with the document identifiers. In either case, the requests

or profiles are updated in accordance with indications furnished to the system concerning the acceptability of items previously retrieved and disseminated. The revised requests or profiles are then used to perform additional searches that may more adequately represent the users' information needs.

The operations of the Smart document retrieval system are briefly summarized in the next section to serve as background for the automatic text processing methods introduced in later chapters.

1-4 AUTOMATIC CONTENT ANALYSIS

The Smart automatic document retrieval system, originally designed at Harvard between 1962 and 1965, is now operating at Harvard and Cornell Universities on IBM 7094 and IBM 360 computers. Smart is a fully automatic text processing system which manipulates documents and search requests, expressed in the natural language, and produces as answers to the search requests the documents that appear to be most similar to the requests. The system is characterized by the fact that

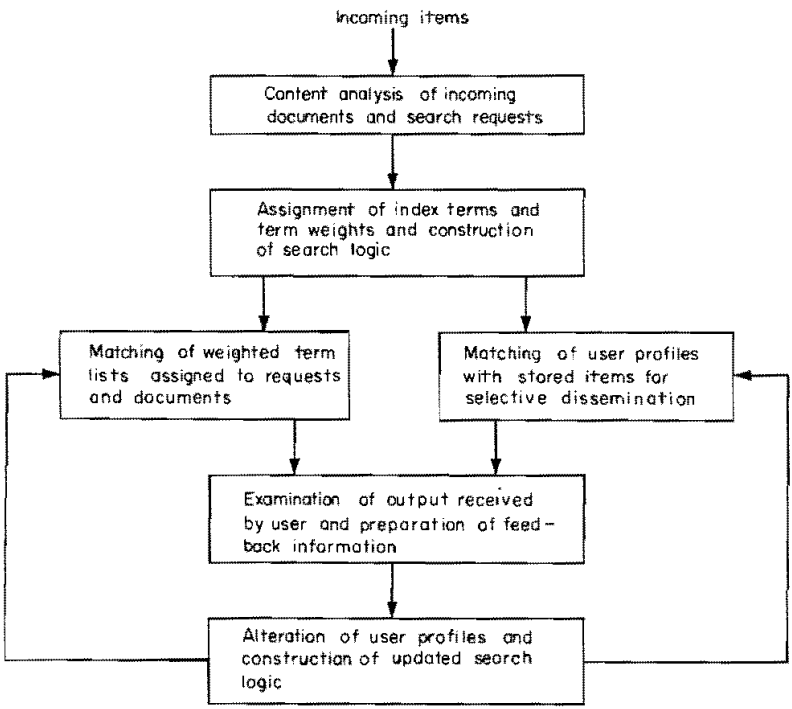


Fig. 1-3 Simplified user feedback process.

several hundred different content analysis procedures are available to generate identifiers for documents and requests, including word matching methods, stored dictionaries to lessen the effect of vocabulary variations, statistical and syntactic procedures to identify relations between words and concepts, and phrase generating methods. The system thus provides the means for attacking the content analysis process from a number of different viewpoints, each producing a somewhat different output. As a result, the search process can be conducted in such a way that search requests producing unsatisfactory results are reprocessed under somewhat altered conditions. The new output can be examined, and, depending on requirements, further changes can be made until such time as the right kind and amount of information are retrieved [12, 13, 14].

Smart is thus designed to serve as a test bed for various automatic information analysis and search procedures that may eventually be incorporated into automatic document retrieval systems. The following summary of its characteristics may be of principal interest in this connection:

1. The information analysis operations incorporated into the system are believed to be sufficiently deep and refined to ensure the identification of much of the relevant material in answer to most search requests.
2. The varying needs of individual users are recognized by enabling each user to call on many different text processing modes and, by choosing a suitable sequence of procedures, eventually to obtain satisfactory retrieval performance.
3. The system also serves as a means for evaluating the effectiveness of a large variety of automatic analysis procedures, in that the same search requests can be processed against the same document collection in many different ways and results compared.

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

1. A system for separating English words into stems and affixes which can be used to reduce incoming texts into *word stem* form
2. A synonym dictionary, or thesaurus, used to replace significant word stems by *concept numbers*, each concept representing a class of related word stems
3. 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

4. *Statistical association* methods used to compute similarity coefficients between words, word stems, or concepts, based on co-occurrence patterns between these entities in the sentences of a document, or in the documents of a collection, so that associated items can then serve as content identifiers in addition to the original ones
5. *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
6. *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
7. *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.

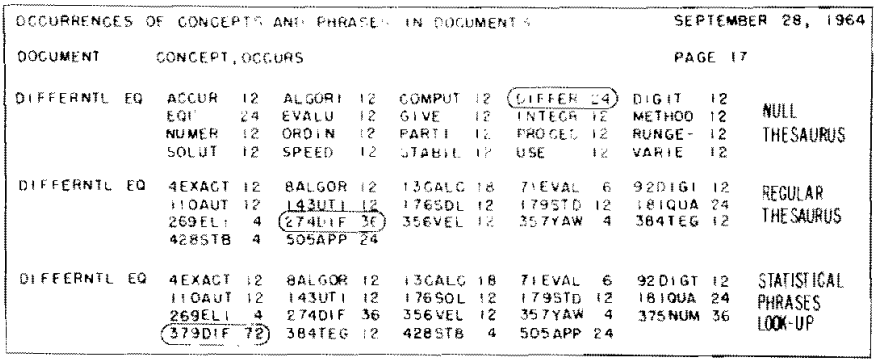
Stored documents and search requests are processed by the system *without any prior manual analysis*, using one of several hundred automatic content analysis methods, and the documents that most nearly match a given search request are identified. Specifically, a correlation coefficient is computed to indicate the degree of similarity between each document and each search request, and documents are then ranked in decreasing order of the correlation coefficient. One or more cutoff points can then be selected, and documents above the chosen cutoff can be withdrawn from the file and turned over to the user as answers to the search request.

As an example of a typical automatic analysis output, consider the search request illustrated in Fig. 1-4. Three automatically produced

```
                                                        PAGE 345
ENGLISH TEXT PROVIDED FOR DOCUMENT DIFFERNTL EQ    SEPT. 28, 1964

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

Fig. 1-4 Typical search request.



OCCURRENCES OF CONCEPTS AND PHRASES IN DOCUMENTS									SEPTEMBER 28, 1964
DOCUMENT	CONCEPT, OCCURS								PAGE 17
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	PROGEL 12	RUNGE- 12				
	SOLUT 12	SPEED 12	STABIL 12	USE 12	VARIE 12				
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							
DIFFERNTL EQ	4EXACT 12	8ALGOR 12	13CALC 18	71EVAL 6	92DIGT 12				STATISTICAL PHRASES 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					

Fig. 1-5 Indexing products for "differential equations."

analyzed forms of this search request are shown in Fig. 1-5. The first consists of a set of weighted word stems; the second consists of thesaurus classes or concepts obtained by a dictionary look-up process (termed *regular thesaurus*); and the last one includes phrase identifiers generated from the *statistical phrase* dictionary. It may be noted in Fig. 1-5 that the weight of the original word *differential* (from "differential equations") was increased by the automatic process from an initial 24, for the word stem process, to 36 for the thesaurus and finally to 72 for the phrase procedure. This illustrates the fact that the automatic analysis can be successful in promoting selected content identifiers that are judged to be important as indicators of document content.

The results of a search performed with the Smart system appear as a ranked list of document citations in decreasing correlation order with the search request, as seen in the example of Fig. 1-6. The output of Fig. 1-6 is in a form suitable for communication with the user who originally submitted the search request. In addition to the list of bibliographic citations contained in Fig. 1-6, special output may also be obtained for evaluation purposes. Table 1-2 shows an example of this output for the search request of Fig. 1-4, using two of the analysis methods illustrated in Fig. 1-5. On the left side of Table 1-2, the 15 document numbers that exhibit the highest correlation with the search request are given in each case, together with the correlation coefficients. The documents that were found to be relevant to the request (as determined manually by an examination of the collection) are marked with an X.

On the right side of Table 1-2, *all* the relevant documents are listed, together with the ranks in decreasing correlation order assigned by the automatic process. Clearly a perfect retrieval system would show ranks

1 to 16 for 16 relevant documents; instead, the regular thesaurus process lists the lowest scoring relevant document with rank 40, and the more powerful phrase procedure produces the lowest ranking relevant document with rank 25. By using the ranks of the relevant documents actually obtained, it is possible to compute evaluation coefficients to measure the effectiveness of each search.

The evaluation measures actually used in the Smart system are based on the standard *recall* and *precision* measures; recall is defined as the proportion of relevant matter retrieved, whereas precision is the proportion of retrieved material actually relevaut. In an operational situation, where information needs may vary from user to user, some customers may require high recall, that is, the retrieval of almost everything that is likely to be of interest, whereas others may prefer high precision, that is, the rejection of everything likely to be useless. Everything else being equal, a perfect system is one that exhibits both a high recall and a high precision [15, 16].

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, it is possible to compute recall and precision values for varying numbers of retrieved documents and to exhibit a plot showing recall in relation to precision. An example of such a graph is shown in Fig. 1-7 for a particular query (number Q145) processed with a collection of 200

```
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
------          -----------          --------------
304STABILITY     0.6675        STABILITY OF NUMERICAL SOLUTION OF DIFFERENTIAL EQUATIONS
                               W. E. MILNE AND R. R. REYNOLDS (OREGON STATE COLLEGE)
                               J. ASSOC. FOR COMPUTING MACH. VOL 6 PP 196-203 (APRIL, 1959)

ANSWER          CORRELATION          IDENTIFICATION
------          -----------          --------------
36CSIMULATING    0.5758        SIMULATING SECOND-ORDER EQUATIONS
                               D. G. CHADWICK (UTAH STATE UNIV.)
                               ELECTRONICS VOL 32 P 64 (MARCH 6, 1959)

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

Fig. 1-6 Output excerpt from Smart system (regular thesaurus process).

Table 1-2 Evaluation of request "differential equations"—15 relevant items

Type of analysis	*Top 15 documents*			*Relevant documents*		
	Rank	*Doc. no.*	*Correlation*	*Rank*	*Doc. no.*	*Correlation*
Regular thesaurus	1	X 384	0.6676	1	384	0.6676
	2	X 360	0.5758	2	360	0.5758
	3	X 200	0.5664	3	200	0.5664
	4	X 392	0.5508	4	392	0.5508
	5	X 386	0.5484	5	386	0.5484
	6	X 103	0.5445	6	103	0.5445
	7	X 85	0.4511	7	85	0.4511
	8	192	0.4106	9	102	0.3987
	9	X 102	0.3987	10	358	0.3986
	10	X 358	0.3986	11	387	0.3968
	11	X 387	0.3968	12	202	0.3907
	12	X 202	0.3907	15	251	0.3329
	13	229	0.3506	17	253	0.3152
	14	88	0.3452	23	390	0.2866
	15	X 251	0.3329	24	388	0.2788
				40	385	0.2301
Statistical phrase search	1	X 384	0.8576	1	384	0.8576
	2	X 360	0.7741	2	360	0.7741
	3	X 386	0.7408	3	386	0.7408
	4	X 392	0.6571	4	392	0.6571
	5	X 200	0.6444	5	200	0.6444
	6	X 85	0.6372	6	85	0.6372
	7	X 387	0.6072	7	387	0.6072
	8	X 103	0.5875	8	103	0.5875
	9	X 102	0.5648	9	102	0.5648
	10	X 390	0.5448	10	390	0.5448
	11	X 358	0.5437	11	358	0.5437
	12	X 388	0.5318	12	388	0.5318
	13	X 202	0.5163	13	202	0.5163
	14	X 385	0.4942	14	385	0.4942
	15	169	0.4794	21	251	0.3444
				25	253	0.3157

documents in aerodynamics. A total of 12 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. 1-7*a*. For the retrieval process illustrated in Fig. 1-7, 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 cutoff 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. 1-7*b*. The recall-precision graph obtained from this table is represented in Fig. 1-7*c*.

To evaluate the effectiveness of the various processing methods used, it is customary in the Smart system to compare output obtained from a variety of different runs. This is achieved by processing the *same* search requests with the *same* document collections several times and making selected changes in the analysis procedures between runs. The differences in the *average* performance of the search requests under different processing conditions are then used to determine the relative effectiveness of the various analysis methods. In the remainder of this volume,

Relevant documents

Rank	Number	Correlation
1	80	0.5084
2	102	0.4418
3	81	0.4212
10	82	0.2843
11	193	0.2731
14	83	0.2631
15	87	0.2594
20	88	0.2315
40	86	0.1856
50	109	0.1631
69	84	0.1305
78	85	0.1193

(*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

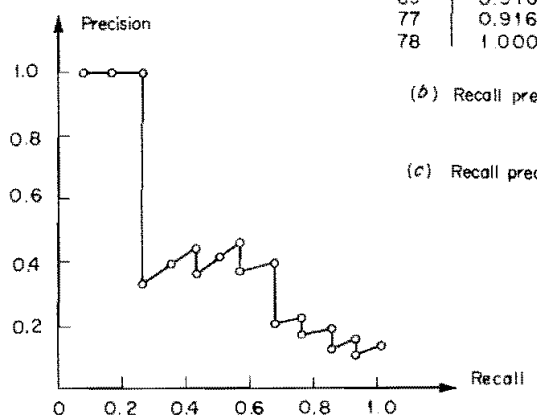


Fig. 1-7 Performance characteristics for query Q145.

Table 1-3 Overall evaluation results. (Based on experiments with four collections in three topic areas)

1. *Term weights*		
Weighted word stems	≫†	Logical stems
Weighted synonym classes	≫	Logical synonym classes
2. *Document length*		
Full summaries (2000 words)	>	Abstracts (150 words)
Abstracts (150 words)	≫	Titles only
3. *Synonym recognition*		
Abstracts with thesaurus	≫	Abstracts word stem process
Summaries with thesaurus	>	Summaries word stem process
4. *Phrase recognition*		
Synonym and phrase recognition	>	Synonym recognition (thesaurus) only
5. *Syntactic analysis*		
Syntactic analysis with thesaurus	≫	Word stem match
Syntactic analysis with thesaurus	>	Synonym recognition (thesaurus)
Syntactic analysis with thesaurus	∼	Statistical phrase recognition (thesaurus)
6. *Term-term associations*		
Stem-stem associations	>	Simple word stems
Concept-concept (thesaurus class) associations	∼	Synonym recognition (thesaurus)
7. *Manual indexing*		
Abstract stem matching	∼	Index term match
Index term with thesaurus	>	Abstracts with thesaurus

† ≫: much greater than
>: greater than
∼: about equal to

averaged recall-precision graphs superimposed for various processing methods are used to reflect the relative improvements obtained from one method to another.

Table 1-3 shows in summary form the main evaluation results obtained up to now with the automatic Smart process, using four different document collections in three different topic areas. The indications are that full document summaries should be analyzed, rather than only titles; that terms assigned to documents should be weighted; and that synonym dictionaries should be used, possibly in conjunction with phrase procedures. Detailed experimental results for a variety of search techniques are included in Chap. 8.

The sample evaluation output of Fig. 1-7 shows that perfect retrieval should not be expected from a single search operation. Experiments

have, therefore, been conducted using a variety of iterative search procedures, where the results of a first search are used to change some of the search parameters to obtain better performance on subsequent passes.

Several possible strategies may be used, including a simple thesaurus display method which enables a user to pick new terms not previously used in order to reformulate his original request, as well as more sophisticated methods using information extracted from previously retrieved documents found to be relevant by a given user. Table 1-4 shows an example where a search request was reformulated by using terms obtained from previously retrieved documents. The improvement in search effectiveness can be ascertained by looking at the ranks of the relevant documents for both the original and modified queries. Table 1-5 illustrates a related process where useless high frequency terms are deleted in the reformulated request and important low frequency terms are reinforced.

An automatic request modification process, known as *relevance feedback*, has also been used in conjunction with the Smart system [17, 18]. In that system, the user is shown some preliminary output and identifies some of the documents as either useful to him or not useful. The system then automatically adjusts the search request by increasing the weight of the request terms that were also contained in the designated set of relevant documents; at the same time, the weight of request terms also contained in the nonrelevant document set is decreased. Effectively, this process "shifts" the request vector so that it lies closer to the relevant document

Table 1-4 Query modification using terms from relevant documents

Query statement (high frequency)	*Terms contained in retrieved documents*	*Ranks of relevant documents*
Original query: "*Automatic* information retrieval and *machine* indexing"	*coordinate, look-up,* search, consult, *abstract,* article, catalog, copy, noun, sentence, science	5, 6, 9, 11, 12, 69
Modified query: "Information retrieval. Document retrieval. Coordinate indexing. Dictionary look-up for language processing. Indexing and abstracting of texts."		1, 4, 6, 9, 11, 18

Table 1-5 Query modification using frequency criterion

Query statement (*high frequency and* **low frequency** *terms*)	*Ranks of relevant documents*
Original query: "Can hand-sent **Morse** code be transcribed *automatically* into English? What *programs* exist to read **Morse** code?"	7, 30
Modification 1: "Can hand-sent **Morse** code be transcribed into English? Recognition of manual **Morse** code."	4, 8
Modification 2: Original query and add **"Morse, Morse, Morse."**	4, 16

subset than to the nonrelevant one. Some typical iterative evaluation output is shown in the recall-precision graph of Fig. 1-8 which contains averaged curves over 17 search requests. The great improvements in recall and precision, particularly between the initially available search requests and the first iterative step, are reflected in the output of Fig. 1-8.

A last point to be mentioned in connection with the design of automatic information systems is the question of the actual strategies to be used for matching the documents with the search requests. Clearly, in practice it is not possible to match each analyzed document with each analyzed search request because the time consumed by such an operation would be excessive. Various solutions have been proposed to reduce the number of needed comparisons between information items and requests. A particularly promising one generates groups of related documents, using an automatic document matching procedure. A representative document *group vector* is then chosen for each document group, and a search request is initially checked against all the group vectors only. Thereafter, the request is checked against only those individual documents whose group vector shows a high score with the request. This two-level search can be extended to a multilevel search by grouping the group vectors themselves and then grouping the groups of group vectors, and so on, as will be seen in greater detail in Chap. 7.

Table 1-6 shows typical search results for a collection of about 500 documents, grouped into 20, 30, and 40 different groups. Each of 24 search requests is first checked against the group vectors and then against all documents contained in the five highest scoring groups. The *group*

Table 1-6 Average results of two-level search for 24 search requests

Total number of groups	*Cumulative number of elements for five groups*	*Total number of comparisons made*	*Group recall*	*Normal recall for full search (same number of retrieved documents)*
20	116	20 + 116 = 136	0.91	0.98
30	82	30 + 82 = 112	0.86	0.97
40	61	40 + 61 = 101	0.82	0.94

recall obtained by the reduced search process is then compared with the normal recall obtained by a search of the full document collection. It may be seen that a search of only a quarter of the collection produces very little decrease in recall, although the results become progressively worse as fewer and fewer documents are used in the search process [18].

The remainder of this volume is devoted to the examination of the principal procedures likely to be of importance in the implementation of information systems. These topics are covered without using any par-

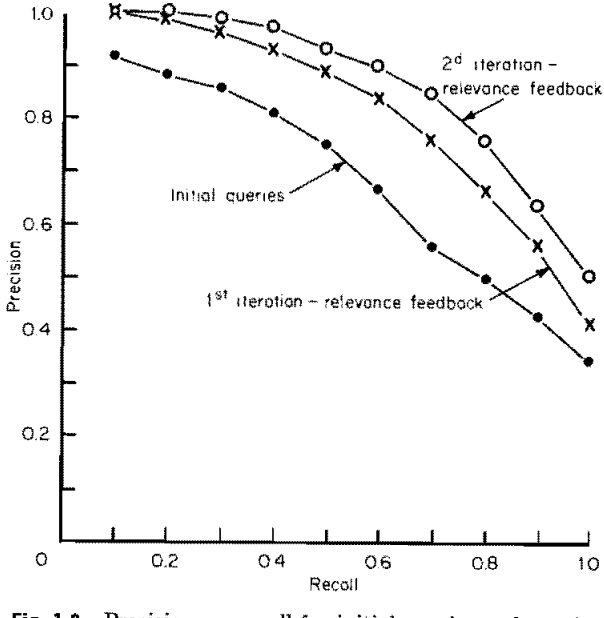


Fig. 1-8 Precision vs. recall for initial queries and queries modified by relevance feedback (averaged over 17 search requests).

ticular existing system or without specialized operations of only restricted applicability. In some cases, results obtained with the experimental Smart system are used for illustrative purposes. The operating procedures used with this system and some typical output products are presented in detail in Appendix A.

REFERENCES

1. Overhage, C. F. J.: Science Libraries: Prospects and Problems, *Science*, vol. 155, no. 3764, Feb. 17, 1967.
2. Shera, J. H.: Librarians against Machines, *Science*, vol. 156, no. 3776, May 12, 1967.
3. Lipetz, B. A.: Information Storage and Retrieval, *Sci. Am.*, vol. 215, no. 3, September, 1966.
4. President's Science Advisory Committee: "Science, Government, and Information," January, 1963. (Weinberg Report.)
5. Carter, L. F., et al.: "National Document Handling Systems for Science and Technology," John Wiley & Sons, Inc., New York, 1967.
6. Rubinoff, M. (ed.): "Toward a National Information System," Spartan Books, Inc., Washington, D.C., 1965.
7. Mikhailov, A. I., A. J. Cherniy, and R. S. Gilyarevskiy: Questions in Scientific Information Theory, *Nauchn. Tekhn. Inform.*, no. 12, pp. 35–39, 1966.
8. Simpson, G. S., Jr., and C. Flanagan: Information Centers and Services, "Annual Review of Science and Technology," in C. Cuadra (ed.), vol. 1, chap. XII, Interscience Publishers, Inc., New York, 1966.
9. Licklider, J. C.: "Libraries of the Future," The M.I.T. Press, Cambridge, Mass., 1965.
10. American Library Association: The Library and Information Networks of the Future, *Rept. RADC-TDR* 62-614, April 8, 1963.
11. Cahn, J. N.: A System of Information Systems, Fourth Institute on Information Storage and Retrieval, American University, Washington, D.C., February, 1962.
12. Salton, G.: A Document Retrieval System for Man-Machine Interaction, *Proc. 19th ACM Natl. Conf., Philadelphia, Pa.*, 1964.
13. Salton, G., and M. E. Lesk: The SMART Automatic Document Retrieval System—An Illustration, *Commun. ACM*, vol. 8, no. 6, June, 1965.
14. Salton, G.: Progress in Automatic Information Retrieval, *IEEE Spectrum*, vol. 2, no. 8, August, 1965.
15. Cleverdon, C. W.: The Testing of Index Language Devices, *Aslib Proc.*, vol. 15, no. 4, April, 1963.
16. Salton, G.: The Evaluation of Automatic Retrieval Procedures—Selected Test Results Using the SMART System, *Am. Doc.*, vol. 16, no. 3, July, 1965.
17. Rocchio, J. J., and G. Salton: Information Search Optimization and Iterative Retrieval Techniques, *Proc. Fall Joint Computer Conf., Las Vegas*, November, 1965.
18. Rocchio, J. J., Jr.: Document Retrieval Systems—Optimization and Evaluation, doctoral thesis, Harvard University; Report ISR-10 to National Science Foundation, Harvard Computation Laboratory, March, 1966.