Automation of Information Retrieval

J. W. PERRY M. M. BERRY F. U. LUEHRS, JR. ALLEN KENT

INFORMATION retrieval is a task that must be accomplished whenever records have to be consulted to provide needed facts. The task may be simple and easy as, for example, in consulting the telephone directory. In contrast, it may require many hours of patient persistence in a well-stocked library to identify and to locate those publications, patents, reports, and similar records that contain information needed in arriving at decisions in various fields of professional activity. This situation frequently arises, for example, when an electrical engineer is confronted by a design problem, when a lawyer must advise a client on a point in law, when a physician must decide on a course of treatment, or when a chemist must develop a process for a needed product.

The role of recorded knowledge in conducting research and development is particularly important to our industrial civilization, as illustrated by Fig. 1. Cost-conscious research management is becoming increasingly aware that successful research is not to be equated with the performance of experiments or the filling of laboratory notebooks with data, however accurate. Rather, research is the application of specialized knowledge to solve well-selected problems.

Information retrieval, considered from the point of view of purpose, enables written documents and similar graphic records to serve as an extension of human memory. To the extent that information retrieval fails to function, libraries and similar accumulations of records become morgues for accumulating embalmed knowledge. The tendency for the accumulation of knowledge to result in its becoming less and less readily available may be counteracted by applying a wide variety of techniques ranging from conventional filing, classifying, and indexing methods through various manual methods, e.g., hand-sorted punched cards, to application of more or less complex automatic devices. Which tools or methods are most appropriate for a given situation must be determined by careful evaluation of such controlling parameters as number of records already accumulated, accession rate of incoming records, rate of obsolescence of accumulated records, complexity of subject matter, breadth of field involved, purposes that the record collection is serving or should serve, present status of organization of the record collection, frequency of use, and urgency for promptness in providing needed information.

It is axiomatic that a system for making information available should be designed so as to provide an optimum margin of advantage in terms of benefits achieved and costs incurred. At the present time, estimation of benefits to be anticipated in a given situation from the automation of information retrieval is no easy matter. As experience continues to accumulate, it is hoped and anticipated that it may prove possible to treat the design of information retrieval systems as an engineering problem in much the same fashion as the design of a solvent recovery system in chemical engineering.[1] However, for the present, the design of digital equipment for automation of information retrieval must be based on rather general considerations.

Acquisition, Recording, Retrieval, and Use of Information

An information retrieval system, by its very nature, is linked, on the one hand, to the acquisition and recording of information and, on the other hand, to the use of information of pertinent interest in a given situation.

As is indicated in Fig. 2, the acquisition of information by observation and experiment either immediately involves or leads directly to a more or less thorough interpretation of observations in terms of existing concepts. In this way, a process of abstraction occurs, to a considerable extent at least, during the acquisition and recording of new factual information. For this reason, any description of an experiment is almost certainly incomplete. Certain observations may not have been made or they may have been left unrecorded as inessential or unimportant. Observations may have been interpreted and recorded in a form that leaves questions in the minds of later workers in the field. For such reasons, a deliberate repetition of an earlier experiment may sometimes be advisable. Even in such cases, however, the repeated experiment can be more expeditiously planned and conducted if the record of previous work is available.

The process of discerning and establishing relationships between phenomena, directly observed, and concepts of a basic, abstract, or general nature continues during subsequent preparation of reports, patents, and similar records and also during their abstracting and indexing. In this way, observational results are interpreted, analyzed, and recorded in terms of concepts which may vary from highly specific, e.g., an individual chemical substance, to highly general, e.g., energy.

The interpretation of original observations may involve considerable mathematical computation. Applications of digital computers to expedite or to facilitate such processing of data are outside the scope of this paper. The realm of activity with which we are concerned is the identification of documents and records that contain information of pertinent interest to a given problem.

Effective use of recorded information that has been identified as pertinent usually requires that it be processed in some fashion or other. Such further processing may be reducible, in some instances at least, to a set of routine operations that could be performed by machine operations. Situations of this type are encountered in banking, merchandising, actuarial, and similar operations.

It seems likely that the future will bring increasing extension of the automatic processing of information identified as pertinent. For example, it may prove possible in the field of tax law to employ automatic equipment to accomplish two purposes. The first would be identification of those statutes, regulations, rules, and court decisions pertinent to a given client's case. The second would be the application of the pertinent law to compute the minimum tax responsibility. Our discussion here is centered on the first type of purpose, namely, the identification of pertinent information. We shall be concerned, in short, with "memory machines" rather than "thinking machines."

Information Retrieval and Machine Capabilities

It is axiomatic that digital equipment can be designed to accomplish any well-defined arithmetic or logical operation or any routine built up from such operations. In theory, it might be possible, perhaps, to design a machine that could scan printed matter or examine other graphic records and then select records of pertinent interest without the need for

J. W. PERRY, M. M. BERRY, F. U. LUEHRS, JR., and ALLEN KENT are with the Battelle Memorial Institute, Columbus, Ohio.

a human expert to conduct a preliminary analysis of the recorded information. A machine for scanning and selecting printed documents would have to be able to perform a series of functions. First, it would have to recognize words, phrases, special nomenclature, and such symbols as the structural formulas of organic chemists and the wiring diagrams of electrical engineers. Recognition of words, phrases, and other symbols by the machine would have to be followed by their interpretation in terms of the information requirements to be met. Such interpretation would require both an extensive memory unit to make the meaning of words, phrases, and symbols available to the machine and also a logical processing unit to test and to detect relationships between the semantic elements that characterize the recorded information being scanned, on the one hand, and the information requirements to be met, on the other hand. A machine able to scan unanalyzed documents and similar graphic records would certainly not be a small digital computer whose modest cost should open up a wide market as an aid to searching and correlating accumulated files of recorded information.

The heart of the problem is how to define tasks to be performed by the digital equipment so that the mechanized searching and selecting system will provide optimum advantages. Particularly careful attention must, of course, be devoted to keeping costs down.

As shown in Fig. 3, equipment cost is not the only item of expense. In an operational system, a considerable investment must be made in analyzing an extensive file of information preparatory to machine searching. Methods for conducting such analysis must be designed so that costs involved in processing documents and similar records do not reach excessive levels. Simultaneously, the design of the equipment must be kept simple in order to avoid excessive construction, maintenance, and operating costs. To meet these two requirements, the same basic principles must apply both to machine design and to the analysis of information.

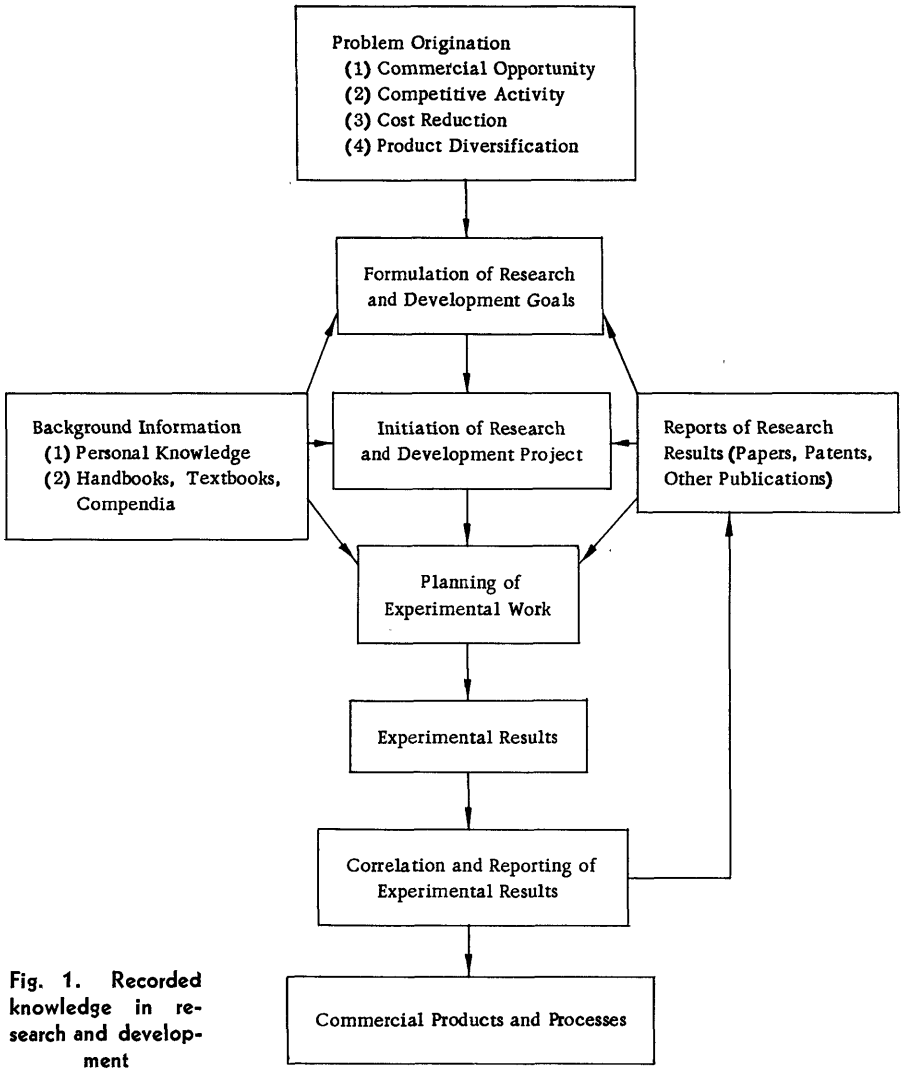

Fig. 1. Recorded knowledge in research and development

Information Retrieval and Class Definition

In seeking a common denominator of basic principles, it is instructive to recall to mind, first of all, that the recording and reporting of factual information in various fields of learning is based on concepts that range from the highly specific, e.g., 2–4–6 trinitrophenol, haematoma, manslaughter, to the broadly generic, e.g., aromatic compound, trauma, crime. Terminology plays a dominating role in recording, abstracting, indexing, and classifying recorded factual information. In seeking needed information, chemists, physicians, lawyers, and other professional men use the same terminology to formulate their information requirements. Such definitions of information requirements will be found on examination, in the majority of instances at least, to involve several terms that are used to characterize the type and scope of information required.

Terminology used to designate various concepts (spatio-temporal entities, abstract concepts, attributes, processes, conditions) may be used to identify one important type of criteria that characterizes both the subject content of records and also the scope of information requirements. It is also important not to overlook a second type of criteria, namely, the relationships between concepts denoted by terminology. Such relationships, in ordinary writing, are denoted by sentence structure or with the aid of prepositions or similar connectives. To make this second type of criteria available as reference points in defining searching operations, it is necessary that important relationships be identified, defined, and explicitly recorded when analyzing information preparatory to machine searching.[2]

The identification and selection of specific records in terms of characteristic criteria may be formulated as the definition of a class whose members are characterized by the specified criteria. The identification of records of pertinent interest may be achieved by matching the characteristics of the subject matter of the records with the characteristics that define the scope of a given information requirement. The matching operations, when formulated on the basis of class definition, provide guidance both for the design of searching machines and also for carrying through the analysis of information.

It should be recalled, in this connection, that the theory of class definition imposes no restriction on the character-

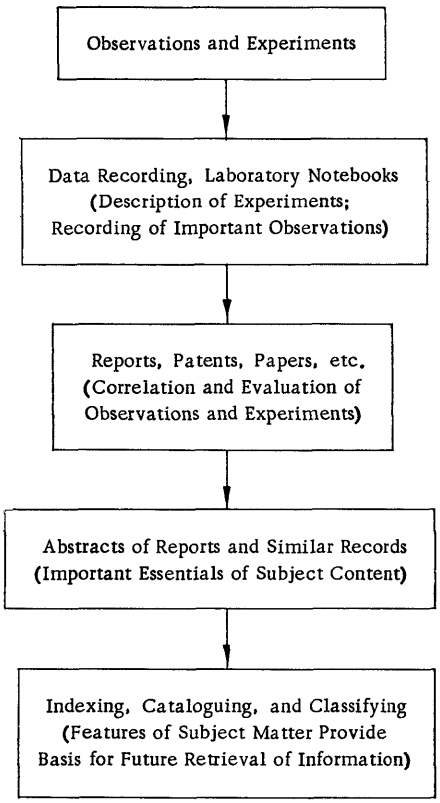

Fig. 2. Steps in acquisition, recording, and processing of information

istics that may be taken into account or, equally important, disregarded when defining a class. The theory of class definition also recognizes the possibility of one class being a member of another class. Furthermore, a class is no less validly a class if it contains only one member or even none at all.

It is instructive to consider how the theory of class definition might be applied to selecting cards on each of which an abstract from "Chemical Abstracts" has been recorded. Perhaps the simplest possibility would be to select those cards that bear an abstract containing some predetermined symbol such as the letter "t" or the numeral "5." Or it might be specified that the abstract shall contain a certain sequence of symbols as exemplified by "nitro," "thermo," "oxidation," or "CH_3." A closely related possibility would be to specify some interrupted sequence of symbols such as "d—tion" or "C_2H_5–H." Such criteria, involving single symbols or sequences of symbols, could be used to detect the presence of certain key words or formulas within an abstract. It is important, of course, to specify that false sequences would not be generated by the interaction of the last few symbols in a word or formula with the first few symbols in the next word or formula scanned by the machine. This means specifically that the machine must be designed so as to detect and respond to certain runs of symbols. One type of run is exemplified by words in ordinary language or by chemical formulas. The next higher order of run would consist of a number of successive words, formulas, or similar symbol sequences to which meaning is attributed. This type of run would correspond to a phrase or clause in ordinary writing. The next higher order of run might be regarded as similar in nature to sentences. Further, still higher orders of runs might be regarded as corresponding to paragraphs, chapters, or complete messages. Symbolically, a run corresponding to a paragraph might be exemplified as follows:

¶SPWxxxWxxxxxWxxPWxxWxxxxxxWxxx
xWxxxPWxxxSPWxxxxWxxxxWPS¶

where

¶ = paragraph start
S = sentence start
P = phrase start
W = word start
x = any individual symbol, e.g., letter or numeral

Note that the start marks designate both the beginning and end of the various orders of runs. As already pointed out, a class of abstracts may be defined by requiring that a given word or formula in the abstract shall contain some symbol or sequence of symbols. Similarly, it may be necessary to identify an abstract as belonging to a class containing information of pertinent interest by specifying that it shall contain at least one higher order run that is characterized in terms of some specified component. For example, in defining the scope of a search we may need to specify that abstracts of interest shall contain at least one sentence characterized by a certain phrase.

It is perhaps obvious that the effective usefulness of a searching machine would be sharply limited, if it were restricted, when scanning recorded messages, to detecting only one criterion of the type corresponding to a single symbol, e.g., "t" or to a sequence of symbols, e.g., "nitro" or "C_2H_5–H." A multiplicity of such criteria may be required to designate the scope of a search. The magnitude of this multiplicity must be held within reasonable bounds, however, as otherwise the cost of providing a large number of detecting units may make the machine excessively expensive. Furthermore, the time and effort involved in setting up the machine to detect an excessively large multiplicity of such search criteria could exceed reasonable limits. These considerations are of importance when deciding how to express the subject content of records in encoded form for machine searching. Before considering this point further, an additional functional requirement of digital equipment for information retrieval should be considered.

Definition of classes becomes much more flexible and effective in literature retrieval if logically defined configurations of criteria can be used to identify abstracts of pertinent interest. To denote such relationships precisely, capital letters, A, B, C, etc., will be used to designate individual symbols or sequence of symbols that characterize words. The three basic logical relationships may then be specified as follows:

1. Logical product
   All of several criteria are required to be present
   Symbolized as $A \cdot B \cdot C \cdot D \ldots$
2. Logical sum
   Any one or any combination of several criteria is required to be present
   Symbolized as $A+B+C+D \ldots$
3. Logical difference
   At least one criterion is required to be absent
   Symbolized as $A-B$

Definition of classes in a form useful for information retrieval may also involve complex logical relationships, as exemplified by:

$(A+B)\,(C-D)+E$
$(A \cdot B \cdot C)+(E \cdot F \cdot G-H)$
$A(B+E+F)-(H \cdot K)$

A fourth relationship specifies that two or more criteria shall be found arrayed in a given order. If we use $<>$ to symbolize such specification of order, then the requirement that the logical product $A \cdot B \cdot C$ is to involve the criteria in that order might be symbolized by $<A \cdot B \cdot C>$. Specification of order may involve more complex logical relationships as exemplified by

$<(A+B)(C+D)(E-F)>$
$(<A-B>+<C-D>)(E+F+G)$

These logical relationships have been specified and illustrated in terms of criteria designated by capital letters. Such criteria, it will be recalled, are individual symbols or sequences of such symbols. By specifying logically defined configurations of such criteria, classes of words may be defined. Thus, the logical product $A \cdot B \cdot C$ could be used to define words characterized by the presence of the letter "a," the letter sequence "tion," and the interrupted sequence "ox–d." A class of abstracts may then be defined as containing one or more of the words that are characterized by $A \cdot B \cdot C$. This class of

words might be designated by A' and other similarly defined classes of words might be designated by B', C', D', ... This opens up the possibility of defining a class of abstracts containing one or more phrases characterized in terms of component words of the classes A', B', C', D', ... As before, such definition of phrases may be based on specification of combinations of words as expressed by logical product $(A' \cdot B' \cdot C' \cdot D' \ldots)$, logical sum $(A'+B'+C'+D'+\ldots)$, logical difference $(A'-B')$, or more complex logical relationships, as exemplified by $(A'+B')(C'-D')+E'$ or $(A' \cdot B' \cdot C')+(E' \cdot F' \cdot G'-H')$.

Similarly, phrases (symbolized by A'', B'', C'', D''....) may be used to define classes of sentences (symbolized A''', B''', C''', D'''...) and the latter in turn used to define classes of paragraphs (symbolized A'''', B'''', C'''', D'''' ....).

Application of the basic principles of class definition in this way provides a common basis for the functional requirements of the searching machine and for expressing the subject matter of documents in machine searchable form.[3]

In summarizing machine characteristics from the point of view of required functions, it is perhaps obvious that a machine searching setup will consist of two major units.

One of these two units is an appropriate medium for recording the abstracts or similar summaries that are to be searched. The selection of an appropriate medium is an engineering question. In general, any medium capable of recording digital information might be considered as a possibility. Some of the more obvious possible media are magnetic tape, punched cards, Teletype tape, photographic film (either in continuous or discontinuous form).

The other major unit in a machine searching setup is the device which scans and identifies which abstracts or similar summaries refer to records of pertinent interest. This searching device must be able to perform a number of functions that might be summarized as follows:

1. Conversion of patterns used to record symbols into corresponding patterns of pulses.

2. Detection of pulse patterns used to record individual symbols or sequences of symbols. (This is equivalent to detection of search criteria designated by $A,B,C,D,$ — in the preceding discussion.)

3. Detection of the beginning and end of runs of symbols where such runs may be of different order corresponding to words, phrases, sentences, paragraphs, and complete messages in ordinary writing.

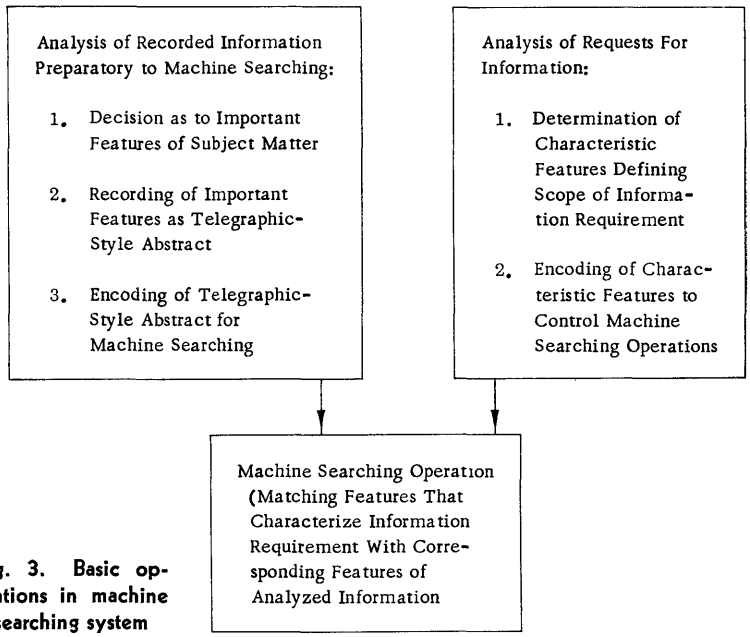

Fig. 3. Basic operations in machine searching system

4. Detection of configurations of logical relationships, either simple or complex, as they may be set up in defining search requirements. Such logical relationships may involve characterizing words in terms of component symbols or sequences of symbols, phrases in terms of component words, sentences in terms of component phrases, paragraphs in terms of component sentences, and abstracts or similar summaries in terms of component paragraphs.

5. Notification of what records have been identified as being of pertinent interest. Such notification may be accomplished, for example, by removing a card from a file. As a recent paper has pointed out, if the "card" is a piece of photographic film, it may bear in microform a readable copy of the abstract or even several pages of printed or other graphic material. Notification of identification may thus be combined with the furnishing of a record of pertinent interest.[4]

These operations can, of course, be performed by a general-purpose computer. A study of the programming required and the effective operational speeds that could be attained supports the conclusion that, for automation of information retrieval on an operational basis, important advantages would be achieved by equipment specially designed and constructed to provide the operational functions summarized.[5]

Analysis and Encoding of Subject Matter

For illustrative purposes, the searching of abstracts as published by "Chemical Abstracts" has been referred to from time to time in previous discussion of class definition as the basis for specifying the functions of digital equipment for searching extensive files of abstracts that relate to complex subject matter. A broad range of flexibility in defining and conducting searches is provided by the afore-specified machine functions. They are, however, obviously inadequate to permit information retrieval to be based in a straightforward and efficient fashion on machine scanning of abstracts or similar summaries written in the English language. The necessity for taking into account the extensive range of synonyms, near-synonyms, and similar alternate terminology available in English and the even wider range of possibilities for alternate phrasing would make it excessively difficult to formulate search requirements so that a high probability of retrieving pertinent information could be assured.

An interpretation of the subject content of documents and records is required to render explicit those aspects of meaning that are to be used as reference points for defining the scope of information requirements and for conducting corresponding machine operations. Such aspects of meaning are of two types. One is ordinarily designated by terminology that denotes spatio-temporal entities, attributes, actions, and processes, conditions, and abstract concepts. When such terminology is used to express the subject content of documents, various important relationships may be involved. For example, it may be important to specify that substance A rather than substance B has a certain property. Such relationships, which are usually designated when writing English by word order, by prepositions, or similar connectives, con-

stitute the second type of important aspects of meaning.

Space limitations prevent a thorough discussion of semantic and systematic considerations involved in rendering explicit, in appropriately coded form, these two types of aspects of meaning. Decisions made in designing a system to meet a given set of requirements must be based on careful consideration of a range of parameters, such as number of documents involved, rate of accession of new material, complexity of subject content, type of information requirement, discriminating ability needed, urgency in meeting information requests, problems involved in distribution of documents identified as pertinent, security regulations, etc.

Searching equipment having functional characteristics as previously discussed provides wide latitude in designing an information retrieval system to meet the exigencies of a given situation. Varying requirements may be taken into account by appropriate decisions as to the following:

1. The degree to which the abstracts, as prepared and encoded for machine searching, provide a more or less detailed analysis of the subject matter of documents or other records. The abstracts may range from brief general annotations to detailed summaries.

2. The extent to which the meaning of words, phrases, and similar terminology is rendered explicit when encoding abstracts. Considerations of cost and convenience make it advisable to record decisions as to the semantic analysis of terminology in a code dictionary. The encoding of abstracts on a routine basis is thereby facilitated. The code dictionary is also helpful in insuring consistency in the analysis of the meaning of new terms and in assigning codes to render their meaning explicit.

3. The type and character of relationships that are taken into account and rendered explicit when encoding abstracts. These relationships may be defined in a broad, general fashion or as specifically and narrowly as may be required by the purposes to be served.

In arriving at decisions, it is necessary to take into account the fact that the discriminating power of a mechanized information retrieval system depends primarily on the extent to which aspects of meaning are rendered explicit by abstracting and encoding. Both generic and specific aspects of meaning can be expected to be important and to require careful consideration. Complexity in a system for analyzing and encoding subject matter inevitably means increased cost, and justification in terms of advantageous benefits must be provided.

Conclusion

Advantageous application of computer-type equipment to automation of information retrieval is determined by ability to analyze the subject content of documents and records in terms of those aspects of meaning that are important in identifying information of pertinent interest. Semantic problems require careful attention in applying automation of information retrieval to meet the requirements of a given situation. In solving these problems in terms of machine operations, the theory of class definition provides the basic framework of reference. Within this framework, there is a wide range of choice as to degree of detail in abstracting and as to rendering aspects of meaning explicit during encoding of abstracts.

Decisions must be made so that automation of information retrieval may provide an optimum margin of advantage as measured in terms of benefits achieved and costs incurred.

References

1. Machine Literature Searching. VIII. Operational Criteria for Designing Information Retrieval Systems, **Allen Kent, M. M. Berry, F. U. Luehrs, Jr., J. W. Perry.** *American Documentation*, Cleveland, Ohio, vol. 6, no. 2, April 1955.

2. New Tools for the Resurrection of Knowledge, Staff Report. *Chemical and Engineering News*, Washington, D. C., vol. 32, 1954, pp. 866-9, 891.

3. Machine Literature Searching. VI. Class Definition and Code Construction, **J. W. Perry, M. M. Berry, Allen Kent.** *American Documentation*, Cleveland, Ohio, vol. 5, 1954, pp. 238-44.

4. The Application of the Kodak Minicard System to Problems of Documentation, **A. W. Tyler, W. L. Myers, J. W. Kuipers.** *Ibid.*, vol. 6, 1955, pp. 18-30.

5. Electronic Digital Machines for High-Speed Information Searching, **P. R. Bagley.** Master of Science Thesis, Massachusetts Institute of Technology, Cambridge, Mass., 1951.

Discussion

W. D. White (Airborne Instruments Laboratory): Could you comment on "Zator" coding of "Uniterms"?

J. W. Perry: It will not be easy for me to comment on matters involving proprietary interests when, in the past at least, the existence of a large measure of disagreement by the proprietors with me has been stated by them on a number of occasions.

There is some uncertainty in my mind whether "Uniterm" is, in effect, a trademark or the name of a specific system. If the latter, then according to what I have read about it, the Uniterm system appears to be based on the philosophy of word-indexing or word-coding. This philosophy might, perhaps, be summarized by saying that the purposes of analysis of information for subsequent retrieval can be achieved by (1) taking the words or, on occasion, phrases found in documents, and (2) using these directly either as index entries or, more generally, as reference points for conducting searches. "You let your material index itself," as this philosophy is sometimes expressed. This seems virtually certain to open the door to all sorts of difficulties with synonyms, near-synonyms, and with alternate modes of expressing ideas in general. The English language's remarkable flexibility, for which Winston Churchill has expressed so much admiration, can only lead to trouble if indexing is attempted without careful control of the use of words and terminology. This conclusion is supported by extensive experience in using conventional alphabetized indexes, particularly poorly constructed indexes. A careful and precise control of terminology, as achieved, for example, by "Chemical Abstracts," is essential to achieving the full measure of effectiveness that is possible with a conventional alphabetized index.

Achieving the full measure of effectiveness with other information retrieval systems also requires that at least equally careful attention be devoted to the meaning of terminology. It is perhaps obvious that the purpose of using automatic equipment to facilitate information retrieval is to expedite accomplishment of the same job for which alphabetized indexing is so widely used. As we tried to make clear in our paper, the essential step in this job is matching the characteristics of an information requirement with the characteristics of the subject content of the documents that are on file. The necessity of accurately defining such characteristics, particularly when they are to be the basis for defining automatic machine operations, is, I am sure, obvious to all. This means that the meaning of words must be accorded very careful attention indeed when indexing subject matter and this general principle holds true regardless of the system, technique, or device used to accomplish retrieval. I stress this point, as I fear there may have been misunderstanding or even disagreement regarding it between myself and the proprietor of Uniterm.

Now to speak of the "Zator" method. Calvin Mooers, proprietor of Zator, is well aware of the necessity of paying careful heed to the meaning—and indeed to the scope of meaning—of terminology. Calvin will not recommend that you use words in a careless or random fashion. He will tell you, and I agree with him completely on this point, that it is necessary to select your terminology very carefully and to use it in a thoughtful fashion if you are going to set up a satisfactory information retrieval system.

Perhaps what I am trying to say is this: If you apply the Uniterm system, taking care, as does Dr. Charles Zerwekh of Hous-

ton, to use terms with close attention to their meaning, then, if the parameters of your information problem are of the right magnitude, the Uniterm system can provide you with advantageous service.

The Zator system is based on the observation that, on the edge of a hand-sorted punched card, or on a similarly edge-notched card, you have a multiplicity of locations. However, this multiplicity is very small by comparison with the number of combinations of locations that can be set up. For example, the number of combinations based on taking 20 locations three at a time is very much larger than 20. Furthermore, it is possible to punch simultaneously on the edge of a card a number of such combinations without having undue trouble with undesirable interactions during sorting operations. "Ghost" combinations resulting from undesirable interactions don't, as experience has shown, become a real nuisance as long as not more than a third of the locations have been punched. Each combination of locations can be ascribed its own meaning corresponding to an index entry or to a characteristic of subject matter of a document. In this way a mechanically performed search can be directed to any one meaningful characteristic or to any combination of the same. But it is perhaps obvious that, here again, your must choose your characteristics carefully and use terminology with precision. If you do this, then I suppose you might consider this an application of Zator coding to the Uniterm system. But, in my opinion, it would be the Zator system straight out.

C. B. Poland (General Electric Company): What study has been made of access times to the data in the computer? In particular, what speed is a probable minimum?

J. W. Perry: Access time and minimum machine speed are variables which depend on other parameters. If the subject matter of your documents is of such nature that they can be divided up into mutually exclusive—or nearly mutually exclusive—categories, then it may be possible to arrange things so that ordinarily an information requirement may be met by searching one or a few categories, in other words, only a fraction of all the documents on file. This can work out in such a way that access time requirements may be met in large measure by simply not having to search a large portion of the file at all. And this, in turn, means that your machine speed can be considerably slower than if you were compelled, for a majority of the information requirements, to search through everything. As may be obvious, the two questions that have just been asked have no simple answers. Rather, the answers to these questions in a given situation must be decided by analyzing requirements to be met, the type of subject matter involved, the number of documents, and similar parameters. Once this analysis has been accomplished, machine requirements can be defined.

The more severe the machine requirements with respect to access time and speed, the higher, in general, you can expect machine costs to be.

P. Bagley of the Massachusetts Institute of Technology (MIT) wrote a master's thesis on this general subject in 1951.[5] If you care to look into these questions in more detail, I'm sure you'll find his thesis interesting and informative.

H. F. DeFrancesco (NSA): Have you performed test scanning on any of the available computers, and if so, please describe the experience and the particular times involved.

J. W. Perry: We were working around to doing this at MIT in connection with the Bagley thesis when, for a variety of reasons, I decided to shift to Battelle. The conclusions in the Bagley thesis were that we would be able to scan, I believe, 5,000,000 documents an hour with a properly designed small computer having the characteristics specified in the paper. The various 5,000,000 documents were considered to have, on an average, 35 index entries, that is, 35 "words" in computer parlance, and it was further assumed that each word would be recorded by not more than 35 digital bits. This is as close as we ever came to an actual trial on an existing computer.

In his thesis, Bagley also investigated the possibility of using the MIT Whirlwind computer to conduct information retrieval along the lines that I outlined in the paper. Programs were worked out, though never actually tested. Even so, there could be little doubt as to what the results would have been. Bagley concluded that effective searching could be accomplished by Whirlwind at about a third the effective rate of an appropriately designed punched-card machine—that is to say, a punched-card searching machine having the operating characteristics outlined in the paper just presented. The reason is that, in a computer such as Whirlwind, you have an arithmetic unit which must make all the decisions. For each decision, data must be fed into the arithmetic unit, the decision made, and the decision placed in storage pending a final over-all decision based on a multiplicity of individual decisions such as establishing the identity of characteristics of information requirements and documents. Before selecting a document as being of pertinent interest a multiplicity of such identifications of characteristics may have to be accomplished, and each identification will almost certainly require a multiplicity of tests for decision by the arithmetic unit. These successive tests constitute a sort of Indian-file parade of information through the arithmetic unit.

This series type of operation turns out to be extremely wasteful in requiring many machine operations to achieve one simple result, e.g., establishing the identity of corresponding characteristics of an information requirement and of the subject content of a document. In particular, many operations, and much time, must be devoted to shuttling information back and forth between the arithmetic unit and the machine's internal storage units. The way out of the dilemma is to design your searching machine—your special-purpose computer—so that identifying operations are performed in parallel rather than in series. This requires establishing a bank of comparator units, each of which may be set to detect some one encoded characteristic of the subject content of the documents being searched. Such identification of characteristics would then be followed by another unit inspecting the decisions made by the comparator units to determine whether the identified characteristics of a document correspond to the logical configuration of characteristics that define the search requirement. The paper just presented reviewed this relationship between identifying operations directed to characteristics and detection of a logical configuration of characteristics. Bagley, in his thesis, pointed out how a plugboard might be used in connection with detecting a logical configuration of characteristics. With this approach, an electronic searching device could be constructed having an effective searching speed several powers of ten greater than attainable with Whirlwind.

Mr. Callen (National Research Council of Canada): What qualifications would you expect in an indexer, and where would you find such people?

J. W. Perry: On talking with various friends in the computer business, they sometimes tell me that the ceiling on the market for computers is determined by the number of available persons having enough mathematical background to do the necessary programming, in particular the interpretation of problems in mathematical terms compatible with computer operation.

In anticipating a similar situation in this field of information searching and correlating, our thinking runs along the following lines. The generation of an encoded abstract for machine searching can be broken up into a series of steps. The first step is someone looking over the document and deciding what is important. This requires a person who has at least a good professional understanding of the subject matter and also a good feeling for the status of the field in question. A person so qualified is in a position to make good decisions as to what is likely to be important, in a given technical paper, to others working in the field to which the paper pertains.

The indexing of the "Chemical Abstracts" is performed by people who have a PhD in chemistry or its equivalent.

Once a person versed in the field to which documents pertain has analyzed them as to important characteristics, the jobs of interpreting these characteristics in appropriate index terms, organizing the index terms into appropriate form, and encoding them, if necessary, should be assigned to persons other than our subject analyst, if costs are to be held to a minimum. It seems virtually certain that problems involving synonyms, near-synonyms, and the meaning of terminology, in general, can be solved by basing the coding of terminology on its semantic analysis. In this way it should prove possible largely to relieve the person analyzing and indexing documents of terminological difficulties. It would be necessary, however, to set up telegraphic-style abstracts to express the subject content of documents. Simple rules for organizing such abstracts are now ready for testing on a pilot-plant scale.

The encoding of properly organized abstracts can be reduced to a clerical routine. For IBM cards, working with recently designed searching equipment, it is convenient to use 3-letter codes to designate basic units of meaning sometimes referred to as semantic factors. Corresponding digital encoding for electronic computer operation would certainly be equally simple to develop.

Your question really goes to the heart of the matter. Permit me to rephrase your question this way: "How do we get large numbers of the right kind of telegraphic-style abstracts prepared?" (Their encoding is, relatively, a much simpler matter.)