THE INFORMATION REVOLUTION

MYTH OR REALITY

G. Salton*

TR 75-240

May 1975

Department of Computer Science
Cornell University
Ithaca, NY 14853

*This study was supported in part by the National Science Foundation under grant No. GJ 43505.

The Information Revolution — Myth or Reality

G. Salton*

Abstract

The early work in the design of automatic information systems, exemplified by the contributions of H.P. Luhn and others, now goes back nearly twenty years. It may be useful to ask whether a great deal has changed in these twenty years, and if so what contributions have been most influential and indicative of future trends. Two contradictory opinions are often heard: the optimists speak of the information revolution and think of mechanized international information networks; the pessimists, on the other hand, assert that little has changed and point to the continued use for well over twenty years of substantially unchanged file organizations, indexing systems, and search techniques.

An attempt is made to separate fact from fiction by describing where we have stood still, and where progress has occurred. In particular, the technological environment which has evolved in many directions is distinguished from most of the intellectual components of information retrieval which have remained unchanged. Suggestions are also made relating to the design and operations of future automatic information systems.

1. Background

Years ago a great deal was written about the information crisis, and dire predictions were made about the information flood which was engulfing the modern

---

Department of Computer Science, Cornell University, Ithaca, New York 14853

This study was supported in part by the National Science Foundation under grant No. GJ 43505.

world. Such sentiments were explained in part by the obvious increase in the production of information items and in part by the political competition between some of the industrial nations. In particular, since knowledge was power, and lack of knowledge consequently spelled weakness, it was imperative to implement bigger and better information services.

This approach to information work, dating back to the nineteen fifties has not totally disappeared in the nineteen seventies. First, a widespread feeling exists that some major problems of the seventies — for example, crime, pollution, traffic control, etc. — cannot be solved without adequate information support. Second, the increasing importance of the "third world" in international affairs brings with it the expectation that effective information services might help the underdeveloped world in "catching up". Finally, some people detect a shift away from industrial, or manufacturing endeavors toward the so-called "knowledge industry":

> "The exponential growth of 'knowledge productions' is not dependent only on growth of population. It is fed by extended educational levels, rapid changes and expansions in technology, and an accelerated shift of the labor force towards the service industries, i.e., universities, governments, and consultants". [1]

Whether or not these arguments are in fact appropriate is immaterial. The interesting point is that one still hears talk about the information crisis, and one wonders whether Bar Hillel's old contentions (that the information problem is mostly a political phenomenon) are still applicable today:

"The geometrical rate of increase in scientific and technological publications ... is commensurate with the increase in scientific and technological manpower ... The so-called problem of information retrieval is therefore ... of a sociological, psychological and, most of all, of a political nature." [2]

Be that as it may, the background — political and otherwise — within which the developments in the information field take place has not fundamentally changed in the last twenty years, even though a decrease in financial support on the part of many governments is noticeable in this area as in many others.

In the remainder of this study an attempt is made to determine to what extent the information processing environment itself may have been transformed in the interim. This is done by examining first the intellectual component and then the technological environment.

2. Intellectual Approaches to Information Retrieval

A) The Punched-Card Heritage

In the early years, information retrieval systems were based on punched-card technology. Typically a number of collectively exhaustive, and hopefully mutually exclusive subject terms, or content indicators, were assigned to the documents of a collection. Two kinds of file organization were available: a <u>direct</u> system in which each card represents a document, the applicable terms being encoded by suitable punches inside, or notches along the edges of the document cards; and an <u>inverted</u> system in which each "feature card" represented one of the terms; the punches then denote documents exhibiting the corresponding subject terms.

In either case, a query formulation would be restricted to include one or more of the existing terms, and retrieval consisted in identifying all documents exhibiting the appropriate assigned term set. In the direct system, needle or mechanical sorting methods could be used to collect all document cards with the proper term set combination. For the inverted file organization, the feature cards corresponding to the requested term set would be collected, and so-called "peck-a-boo" or optical coincidence procedures would produce the documents jointly encoded on all chosen feature cards.

With the punched-card technology, retrieval was an all or nothing affair: a document would be retrieved if it was identified by the appropriate term set, and it would be rejected otherwise, because retrieval was strictly dependent on finding punches in certain specified positions of the cards. No computational process was involved, and thus no approximately relevant items could be found. Neither was there any ranking of the output in some presumed order of importance of the corresponding documents. As Fairthorne states:

> "punched card machines ... (function as) mechanized pigeon-hole systems where requests are restricted to the pigeon-hole headings, and the headings are almost, but not quite, mutually exclusive." [3]

In summary, the punched-card technology provided retrieval based on keyword (feature, heading, term, etc.) indexing, and query formulations consisting of simple keyword sets. The system would treat each query as a conjunction of terms, that is, a collection of terms joined by Boolean "and" connectives; no other query types could be handled.

In comparison with the normal computer environment, a punched-card retrieval system is inexpensive. However, the size of the document set that can be accommodated is necessarily restricted. Furthermore, the effectiveness of the system depends greatly on the choice of the available keyword set, on the accuracy of the indexing operation, and on the size of the retrieved document set corresponding to a given query statement. [4]

The intellectual features of existing, operational computer-based retrieval systems are all directly derived from the earlier punched-card implementations. The main components are described in detail in the next few sections.

B) Simple Keyword Indexing

It is only a slight exaggeration to state that all existing, operational information systems are based on simple keyword indexing, that is, a document is identified by sets of terms, or keywords, used very much like those inherited from the punched card days. Furthermore, almost universally, the keywords are assigned manually by trained indexers or subject experts, and attempts at mechanizing the indexing operation are not generally designed to replace the simple terms for content representation but only to alter the manner in which they are assigned.

In view of the many criticisms and demonstrations of inadequacy, it is surprising that the indexing procedures have now remained unchanged for several dozen years. Years ago, Bar Hillel has stated that:

> "it is a logical category mistake to assume that a word, or a phrase, contains information in the same sense in which a statement does. In spite of prima facie appeal, the information content of a statement is not the sum, or combination, of the information content of its constituent phrases ..." [5]

Bar Hillel then goes on to show that the Boolean conjunction of terms as used in keyword retrieval leads to many erroneous content identifications. Thus, while the set of "red houses" is indeed the set of items that might be identified by the terms "red" and "house", the set of "Newfoundland dogs" is not equivalent to items that are identified by both "Newfoundland" and "dog". [5] In other words, the representation of document content by collections of terms that are juxtaposed for retrieval purposes entails a variety of shortcomings, regardless of whether the terms are assigned by human beings or automatically.

The latter possibility, known as automatic indexing, is itself criticized on additional grounds as shown in the next section.

C) Automatic Text Searching

It is a commonplace observation to note that the words or phrases occurring in the text of a document may, or may not, be representative of document content. Bar Hillel is unequivocal about his feelings in this respect:

> "Could not the stage of (manual) clue assignment be completely skipped and the request topic be directly compared with the original documents ... there is nothing in our knowledge of the workings of communication which would indicate that such a proposal is, or ever will be, practical." [6]

Bar Hillel cites as an example the person wanting documents on "copper plating" when the document contains the sentence "in this paper we shall not deal with copper plating," and he asks how a machine could possibly identify synonymous expressions.

These opinions concerning the nonfeasibility of text matching for indexing purposes were widely shared twenty years ago, and surprisingly, they are still held today:

> "neither the assignment of topic terms to a given request, nor the reformulation of a given request are processes which could conceivably be adequately mechanized, contrary to some speculations in that direction." [6]

or again

> "at this time we feel that human beings perform the necessary and sophisticated task of deciding what is important in an article far more reliably than a machine can be programmed to do." [7]

Bar Hillel, and others postulate a number of requirements that would have to be met before automatic indexing could become a possibility. Among these, the following are of main importance:

a) the existence of a structural arrangement of topic classes — for example in tree or lattice form — showing a variety of important relationships between the included topics;

b) some notion of distance, or similarity, between different topics;

c) the availability of a reasonable coefficient of similarity usable to determine similarities between pairs of distinct documents.

While it would be nice in principle if these requirements could be met in practice, the chances of this happening are very small indeed. A great deal of work has been devoted over the years to the general theory of classification, and a large number of specific classification systems have been built. All of them have been found wanting in various respects, and none of them can be used to obtain a unique distance between topic classes that would hold under all circumstances. Bar Hillel characterizes these efforts in the following terms:

> "the failure of these attempts (at defining distance measures between topics), I would say, are not due to the insufficient stature of those who made them, but rather to the fact that they attempted the (almost) impossible ... no approach being in view that has any chance of leading to a function that will give us a useful measure of distance between topics, and still less between documents" [6]

No formal results are available concerning the impossibility of properly defining topic distances. However, Mooers has shown that the famous triangle inequality (which says that the distance between two items cannot be greater than the combined distances via a third one) does not necessarily hold in a document space. [8] Mooers postulates the existence of two documents X and Y with no common terms, plus a third one Z that contains one half of its terms in common with X and one half in common with Y. The three documents then have the following structure:

$$X = (x, x, x, x, x, x, x, x)$$

$$Y = (y, y, y, y, y, y, y, y)$$

$$Z = (x, x, x, x, y, y, y, y)$$

where the x's and y's represent terms assigned to X and Y respectively. If the similarity coefficient S between two documents is defined as the proportion of matching keywords, we have

$$S(X,Z) = S(Y,Z) = 1/2$$

$$S(X,Y) = 0$$

Normally the distance d between two items is defined as an inverse function of their similarity — that is, the larger the similarity, the smaller will be the distance. It follows that

$$d(X,Y) > d(X,Z) + d(Y,Z)$$

contrary to the normal condition in Euclidian space. [8]

Assuming that documents such as those postulated by Mooers are actually found in practice within a given document collection, one must conclude that effective retrieval becomes impossible, since documents that closely resemble a query may then be retrieved after those that are more distant contrary to normal expectations and requirements.

In summary, the thinking twenty years ago was that single term indexing produces an unwarranted simplification of document content, and that assigning the content identifiers automatically makes the situation worse, because machines are unable to determine relationships and distances between terms and documents.

The questions relating to the use of single term indexing and automatic content analysis are taken up again later in this study when appropriate experimental evidence and some theoretical results obtained over the past twenty years are given.

D) The Boolean Algebra Model

One additional component of existing retrieval systems needs to be mentioned, and that is the universal custom of formulating search requests using Boolean term combinations. A typical query might then be formulated as "(a <u>and</u> b) <u>or</u> [(c <u>and</u> d) <u>and</u> (<u>not</u> e)]" where a, b, c, d, e are suitable index terms. Such a formulation is a direct extension of the punched-card model, and its use has become so ingrained that when an attempt is made to eliminate it in favor of a simple term vector model, the users may be expected to complain:

> "we found it expedient to implement an explicit Boolean capability, with all its well-documented drawbacks; ... this experience taught us that information-seeking behavior is certainly <u>not</u> as succeptible to change as we had initially assumed it to be ..." [10]

Several problems arise when Boolean query formulations are used for retrieval purposes. First is the underlying assumption that when document set A corresponds to some term a, while document set B corresponds to term b, then the query "a <u>and</u> b" corresponds to the intersection of document sets A and B (that is, $A \cap B$). In actual fact, the set operations defined in a Boolean algebra, consisting of set intersection, set union, and set complementation, do not necessarily correspond to the way in which terms in the language are combined into semantically acceptable entities. (It was previously alluded to that the set of items dealing with "fish food" is not the intersection between the items dealing with "fish" and those dealing with "food".) [5]

Second, there are problems with set complementation. In a normal Boolean algebra, the set of items corresponding to "not a" includes all the items to which "a" is not assigned. But linguistically, the complement of a term normally denotes a much smaller set. For example, "not yellow" normally refers to items that are colored in someway that is different from yellow — for example, red, green, blue, etc. But there will be many items for which the concept of color simply does not apply, or for which we do not know what color may apply, if any. Such items would not be included either in the set "yellow", or in the set "not yellow".

A somewhat similar point is the one made by Fairthorne when he remarks that in a growing, dynamic environment, where new items are received periodically that are not immediately indexed, the set complement is not uniquely defined:

> "the system can no longer retrieve all those and only those items called for ...; there are now two types of complements: one is the largest set of items that certainly does not contain any "a's" — the pseudo-complement ...; the other complement is the smallest set of items that certainly contains all the "non-a's" — the Brouwerian complement ..." [11]

Perhaps the most damaging effect of the Boolean environment in a practical retrieval situation is the lack of discrimination among the items indexed by the requested term combination. All such items are normally treated as equally valuable by the system: no relevance weight is assigned to any output item, and no ranking is therefore provided for these documents in presumed order of importance. In these circumstances, the user may be faced with 50, or 100 or more retrieved documents, all of these presented in random order without indication of reasonable processing order. [12]

All these facts are well known to the experts, and they have repeatedly been pointed out in the published literature:

> "it is a common fallacy, underwritten at this date by the investment of several million dollars in a variety of retrieval hardware, that the algebra of George Boole is the appropriate formalism for retrieval system design; this view is as widely and uncritically accepted as it is wrong." [13]

The same facts have emerged more recently as a result of evaluation work performed on various modern operational retrieval systems:

> "The use of Boolean algebra for querying computer-based retrieval systems may have been a mistake ... arising through the similarity of early computer-based retrieval systems to systems employing edge-notched cards, or the optical coincidence principle ..., there was no possibility (in punched-card systems) for partial match and for the retrieval of lists of references ranked according to degree of match with a search query ... Computer-based systems have no reason to be so restricted." [14]

In spite of all these misgivings and exhortations practically all retrieval systems operating in 1975 in a real user environment continue to be based on manually assigned term sets utilized in a Boolean framework for search formulation and retrieval purposes. There has been no basic change in the last several decades.

E) Formal Intermediate Languages

A direct result of the alleged inadequacy of the single-term indexing languages is the flurry of activity aimed at designing formal indexing languages capable of representing document and query content more completely and more accurately. Invariably, such formal languages include dictionaries or authority lists describing the permissible vocabulary, as well as rules for defining and representing relations between terms.

The former, that is, the definition of the permissible vocabulary is of course dependent on the subject area under consideration and is not of particular interest for present purposes. The rules relating to term relations, on the other hand, are important on several accounts: first because so many people claim that an adequate content description without inclusion of term relations is inconceivable:

> "(the introduction of) relations between terms is the central semantic problem of subject indexing;" [15]

second, because the system of relationships that is believed to be necessary and appropriate can in principle be independent of subject area; and finally, because the design of an acceptable and complete system of term relations would in fact constitute a great step forward in the theory of classification and related areas.

A large variety of different systems of relationships between terms has been proposed in the literature, including special "role indicators" to be assigned to each term, and "interfixes" to specify relationships between terms. A few specific formal systems may be cited for illustrative purposes:

> "The researchers found that all things can be reduced to four fundamental types: differentiation, figuration, categorization and correlation. Differentiation ... allows us to speak of warm and cold, of silence and noise, of hard and soft, etc. ... Figuration gives us figures and forms .... Categorization gives us the mental, or logical categories, including, for example, substance, accident, subject, object ... etc. Correlating is giving a temporal order to the material received from the other activities ...." [16]

Another formal system for the representation of information content due to Ranganathan recognizes five fundamental categories, or facets, including "personality", "matter", "energy", "space", and "time". Each facet may be further subdivided into so-called "rounds" and "levels". [17]

Many of these systems have actually been tried out in the sense that the respective formal languages have been incorporated into the design of operating retrieval systems. A typical example is the so-called "semantic code" originating at Western Reserve University. [18] However, most of the proposed systems have proved to be ephemeral, and not a one can really be said to have withstood the test of time. Either the proposed relationships were not as complete and unambiguous as had been believed, or else they could not be applied equally easily to all subject areas, or the proposed indexing process was found to be too elaborate and complex for practical use.

Furthermore, the emphasis on the use of controlled terms and term relationships may have reflected an unwarranted concern with retrieval precision at the expense of recall. Specifically, the experts in retrieval operations appear to have been concerned exclusively with what in the punched-card days used to be called "false drops", that is, the retrieval of extraneous materials. In consequence, attempts were made to render the document descriptions and search formulations extremely precise, implying the use of term relations of various kinds:

> "the basic disadvantage (of most other systems) is the low accuracy of their representation of detailed subject headings ... (this) creates the possibility of many parasitic combinations of characteristics which may represent headings having a quite distant relationship with the original heading ... Consequently the answers will contain a large amount of unnecessary material." [19]

In the process, one may have lost track of the fact that precision and recall vary inversely. In particular, while a very precise search formulation will certainly retrieve few extraneous items it cannot be depended on to retrieve a great many relevant items either. A good deal of recent experimental evidence obtained by Cleverdon and others has shown that specific, complicated item descriptions and search formulations produce disastrously low recall results, and may be expected to reduce overall retrieval effectiveness.

The conclusion must then be the one reached years ago by Luhn when he argued that manual indexing methods using formal intermediate languages could not be expected to produce uniformly effective retrieval results:

"under the circumstances, there is serious doubt as to the effectiveness of storing information in translated form, that is, coded in terms of a common language for machine searching and translation ...." [9]

In recent years many of the efforts devoted to the construction of artificial retrieval languages have thankfully been abandoned, and the emphasis in this area has shifted toward question-answering and fact-retrieval applications.

F) Summary

In summary, so far as the intellectual components of retrieval systems are concerned, the introduction of the computer environment has produced practically no change from the punched-card implementations. There is a continued insistence on the use of manual indexing of information items and manual encoding of search formulations, even though the experimental evidence does not support the contention that manual indexing products furnish better retrieval effectiveness. Furthermore, whereas appropriate linguistic theories continue to be lacking and only imperfect models exist of language structure and understanding, the arguments by Bar Hillel and others concerning the necessity of intellectual input in the content analysis and related operations are still taken at face value:

> "I can find no instance in which the intervening years have proved him (Bar Hillel) wrong ...." [20]

This leads to a situation where on the one hand sophisticated content descriptions are believed essential, while on the other no viable methods are available to provide them.

Somewhat analogous arguments apply to the use of Boolean logic in search formulations and to the Boolean algebra model in general. The misgivings of Bar Hillel's, Fairthorne's, Hillman's, Lancaster's, Mooers' and others continue to be ignored presumably, in part, because alternative methods, such as the vector space model, are not as easily understood. Manual term assignments and the use of Boolean connectives have now been utilized in practice for fifty years.

The situation concerning the organization of the stored information files is similarly unchanged from the punched-card days. Direct files in chronological, or document acquisition order are used when the file sizes are sufficiently small to permit individual comparisons between all stored items and each search request. When rapid responses are required, or when the files are extensive, an inverted file directory is used requiring the listing of a document reference under each applicable index term. This renders it impossible ever to change the term assignment for any document, and limits the index terms available for content representation to the initially chosen static vocabulary. The "peck-a-boo" retrieval strategy developed in the punched-card days is thus carried over to the modern computer environment.

An additional surprising fact: information retrieval managers have long been faced with the multitudinous problems connected with file growth and with the dynamic nature of information files. Nevertheless, they continue to operate systems with static indexing vocabularies, and they have been uninterested in developing file retirement procedures and other measures designed to subdivide large file arrangements into smaller, more efficient, subportions.

Alternative methodologies for the performance of various intellectual retrieval processes are mentioned following an examination of the modern computer environment.

3. The Technological Environment

A) Some Ideas Concerning the Role of the Computer

For most people a computer is a tool that makes possible the rationalization of information processing tasks when applied imaginatively to the problems at hand. It may come as a surprise that such a view is considered simple-minded by some observers. Two additional viewpoints may be heard in particular, that for lack of better terms may be termed the "undersell" and the "oversell". The undersell consists in pointing out that computers are mechanical calculating devices and that their utilization implies a view of the information field which is essentially mechanical and quasi-deterministic:

> "In information science, as well as in other disciplines, the computer has become ... a way of looking at the world. The computer has emerged as a cultural phenomenon. I call it "the gestalt of the computer" .... Broadly speaking, the computer has caused us to view human information processing as analogous to machine processing ... and although I believe that much of man's intellectual and physical behavior is indeed mechanical, I do not believe that it is essentially mechanical." [20]

The argument seems to be that this so-called "world view" of the computer reduces most intellectual processes to those of fundamental logic and to mathematically derived relationships between entities. Since people like Bar Hillel and others have shown that the Boolean model is faulty and that language understanding by computer is impossible, one concludes that computers cannot serve as fundamental components in information science and that their scope and role in the field should be appropriately limited:

> "Information science must abandon its deterministic approach and must recognize the computer as perhaps an historical accident rather than a scientific organizing principle. We must get out from behind the computer ... we must begin to pay more attention to the social, cultural, and spiritual aspects of human communication. We must begin to realize that what a man says or writes is not simply the additive sum of the phonemes or the morphemes." [20]

The latter thought is an imitation of Bar Hillel's, but the total argument is a curious mixture of mysticism and half-truth. The "gestalt" of the computer, whatever this may be, appears to this writer to be immaterial. The utilization of a computer for an information processing task should not imply fundamental alterations in the <u>nature</u> of the task itself, although of course the organization of the process and the particular manner of performing the task may change when a computer is introduced into the environment. There are indeed ways of utilizing computing machines unimaginatively in a quasi-mechanical way; obviously, the role which a computer may play in the performance of a given task depends on the workers that use them. However, there is no evidence at all, that the "social, cultural, and spiritual" aspect of the information field has anything to do with the utilization of computers in a given processing environment.

A different, although in some ways congruent, view of computer capabilities in the information field is that computers do indeed produce fundamental changes in the retrieval situation, not because of their essentially mechanical nature or "gestalt", but on the contrary because of the enhanced intellectual perspective afforded by their use. [21] This envisions a so-called "on-line intellectual community" in which a typical reseacher would conduct his work in a kind of symbiosis with the computer, using a console device, on-line, for continuous communication with the machine:

> "he acquires data in computer-processible form; he couples statistical analysis closely to the acquisition of data ... he may come to the point at which data from the library or from a data bank should be compared with data newly acquired through experimentation ... he is likely to join the laboratory and library part of his work more closely together. [21]

This type of close and continuous interaction with the computer would then produce an optimum merging of the algorithmic and heuristic components of the problem-solving tasks, the former being of course most appropriately handled by the computer, and the latter more specifically a human function:

> "most constructive thought involves a playing, back and forth, between the heuristic and the algorithmic functions." [21]

An extension of the single-user model provides for a whole community of users interacting freely through a computer and/or information network. This allows the automatic system to be used not only for research purposes, but also during the report writing, publishing, and information dissemination functions:

> "There is at least a reasonable hope that on-line information networks will facilitate cooperative interaction in intellectual endeavors ... (They also) introduce the possibility of achieving very close interaction between the substantive scientist or engineer and the documentation expert ... during actual preparation of a report, and continue through the process of editing, publishing, and incorporating into the body of knowledge." [21]

While it would be a mistake to dismiss such an environment as impossible of achievement, it is also a fact that the implementation of flexible computer and information networks is behind schedule compared with the early predictions. Furthermore, a variety of problems relating to the provision of inexpensive communication links between remote locations, the development of input-output devices, and the implementation of flexible storage organizations and search techniques render the proposed environment less immediate than one might expect. It would then be a mistake to restrict information research plans exclusively to those components requiring the close human-computer interaction which is contemplated for the future.

B) Actual Computer Environment

It may be worth inquiring whether the computing capabilities useful in a retrieval environment have indeed changed in fundamental ways over the last two decades, and if so to what extent such changes are likely to produce improvements in retrieval practice. The answer to the first question is a qualified yes. Certainly the flexible type of computer network postulated in the previous section does not exist as yet, but far more powerful machines and far greater memory capacities are available at the present time than was the case twenty years ago. This implies that larger information files can now be processed, and that fast file searches become possible.

The improvements in the computing and storage situation have been matched by advantages in the available input-output equipment and in the input conversion of bibliographic information. Quite sophisticated console terminal devices now exist, including some with graphic display capabilities, nearly noiseless typewriter keyboards, and high resolution text display facilities. Problems still arise in converting text into machine-readable form since versatile automatic text readers are not available. However, the multiplicity of tape distribution services that now disseminate bibliographic information in machine readable form alleviates the input problem to some extent. And while the previously-mentioned communication nets are still too costly for general use, that situation may change in the future.

One possible model for a future information processing capability is typified by the experiments of project Intrex. This project was started with the following aims in mind:

> "the success of time-sharing, interactive computer service in the academic community suggested that similar techniques might be invoked to deal with the library crisis ...; it was not unreasonable to hope that the technology of data transmission and data processing would open the way to more effective storage and retrieval of technical information." [22]

The Intrex environment included an on-line storage and retrieval environment using cathode ray storage and teletpyewriter terminals. The system made available in machine-readable form for user interrogation a variety of catalog files, and the full text of all documents included in the system stored in a microfiche file. In fact, the same graphic console devices made it possible alternately to view document citations stored on the standard magnetic medium, and the full text from the microfiche store.

A good deal of experimentation was performed with the prototype equipment, and the conclusion was reached that the proposed services were technologically feasible, although too costly to render under present conditions. The hope was voiced that in time costs might be reduced, and several grounds were cited for optimism: first was the expectation that the cost of electrical transmission would decrease as the use and availability of cable television (CATV) became more widespread; second were the economies resulting from a general sharing of collections and data bases by a large aggregate of potential users. Obviously, as the acquisition and maintenance of information items becomes more difficult and more expensive, individual libraries and information centers become less capable of satisfying the needs of their user populations. In these circumstances, the use of centralized information facilities with widely dispersed user populations proves increasingly attractive. Some question exists concerning the exact form to be taken by the mechanized information facilities of the future, and the technologies to be used; however, no serious doubt arises concerning the ultimate desirability of collection sharing, on-line information access, and mechanized search and retrieval capabilities.

One particular aspect of the current technological environment may require special comment: this is the appropriateness of on-line technologies for performing many of the potentially desirable information processing tasks, including query formulation or reformulation, browsing through the collection, broadening or narrowing the scope of a query, and so on. A division of opinion arises here caused in part by the varying roles assigned to the computer by different observers in information processing. Bar Hillel who considered this question many years ago was not hopeful about the possibilities of utilizing computers for sophisticated collection searches. Under the heading of "getting something for nothing", he states:

> "to claim that in the system of bibliographic coordination, by combining various terms, we may get information which although implicit in the system, was never explicitly recognized is plain buncombe ...; (in fact) the whole idea of approximating the association chain of ideas in the human brain by a cooccurrence chain of index terms ... seems to hold very little promise." [5]

A large number of other workers, including many that have actually tried interactive searching in operational or in laboratory situations are delighted with the possibilities:

> "in our opinion, on-line interactive searching is not just another efficient and effective information retrieval technique, it has the potential to become a truly innovative search tool ..."
>
> "the basic nature of the interactive process dictates that additional, useful search terms frequently turn up in the process of search construction ... we have also found that "quick and dirty" searching can often be substituted for major search effort ... interactive searching seems to provide a psychological boost to satisfaction with search quality." [23]

In the final analysis, it may not be necessary to establish whether on-line retrieval environments in fact contribute to congenial intellectual situations because the economics of information acquisition and processing may mandate some form of collection-sharing and eventual remote access to the collections for most search and retrieval systems.

C) Current Efforts

The foregoing development indicates that the hardware environment has been subjected to considerable alterations over the last twenty years, in the sense that faster and more powerful machines are now available, and that different types of processes and user system interactions can be carried out. It was seen earlier that this altered hardware environment has caused only minimal changes in file organization and query formulation techniques and that the basic retrieval methodologies have remained constant for many years.

Certain innovations in processing capabilities that may be indicative of future trends should however be noted. These are of three main types: practical search tools and products, incipient centralized or cooperative processing, and on-line retrieval services. These developments are all distinguished by the fact that the intellectual content manipulations are handled by conventional procedures, whereas the simpler file maintenance and search operations are mechanized.

A number of practical search tools have been developed in recent years using automatic or semi-automatic methodologies, some of which have found wide acceptance in practice. Among these are the so-called permuted title indexes, sometimes called KWIC or KWOC indexes, exhibiting references to bibliographic items arranged under the various terms of the corresponding document titles. KWIC indexes provide document retrieval by using title words as main entry points. A second example of an automatically generated search tool is the well-known citation index which lists for each bibliographic entry the set of more recent documents containing citations to the original. Citation indexes lead to the retrieval of documents based on the respective citation patterns; they are also used for historical studies, to trace the evolution of scientific disciplines, and to evaluate the influences of individual authors.

Retrieval aids such as KWIC indexes and citation indexes do not furnish direct access based on actual document content. By using indirect indicators such as citations or title words, the corresponding lists can, however, be produced automatically without requiring complicated content analysis methods. Title word and citation tracing are actively used in many operational environments.

A variety of cooperative information processing arrangements have already been implemented in the library field. Eventually, these may lead to collection sharing and to common library housekeeping methodologies. One of the most influential of these cooperative systems may be the on-line cataloging facility at the Ohio College Library Center (OCLC). The OCLC organization maintains large bibliographic files that are automatically searchable using console devices situated in many remotely located participating libraries. The multiplicity of individually maintained catalog files is then replaced by a single, machine-readable bibliographic file used cooperatively by many organizations for cataloging and for the production of library catalog cards.

Once again, the intellectual content analysis and classification operations are carried out by human operators, but the files are mechanized and many pencil and paper operations are replaced by on-line manipulations of computer console displays.

The last class of ongoing activities requiring discussion are the many retrieval services that are now in operational use in many different subject areas. A great variety of selective information dissemination services provide output of newly arriving materials to users on a continuing basis. Some of these SDI services include automatic user profile (query) updating methods based on information furnished by the users following earlier retrieval activities. While these services normally operate off-line in a batch-processing mode — a large number of user queries being processed simultaneously — they provide perfect environments for the use of customer feedback information to improve future retrieval output.

Unlike the selective dissemination services, an ever-increasing proportion of the on-demand search operations (where user queries are individually submitted before a given search is actually carried out) are now conducted on-line by the users themselves, or by appropriate search intermediaries, using console terminals for user-system communication. The search files typically consist of bibliographic information made available in machine-readable form by one or more of the existing tape distribution services. In some organizations large libraries covering many different subject areas are made available for interactive search and retrieval. [23-24]

Many of the drawbacks of the earlier punched-card or batch processing retrieval implementations are present also in the modern on-line systems: the difficulty of formulating effective query statements when the characteristics of the retrieval vocabulary are not known, the problems arising from the use of Boolean connectives as part of the search statements, the lack of control over the output size, the difficulty of providing ranked output in decreasing order of presumed importance, the inconsistency with which content identifiers are assigned manually to queries and documents in many cases, the static nature of indexing vocabularies and file organizations.

On the other hand, substantial improvements are obvious in the physical, if not necessarily the intellectual retrieval environments: the fact that access can be obtained to much larger and more diverse information files than was previously possible, the increased flexibility in the available search strategies, the opportunity of aborting searches that go wrong and of reformulating inadequate query statements, the possibility of performing rough trial searches in order to acquaint oneself with system and collections, and of course the nearly instantaneous response rate. These improvements account for the optimism voiced by many users of on-line services:

> "We are convinced that on-line interactive searching is a research and information tool that becomes truly effective when it is treated seriously; under proper conditions we believe that the results obtained are challenging and even exciting." [23]

4. New Retrieval Environments

A) Management Aspects

Given the apparent mismatch between the modern on-line technologies that are made available for information processing purposes and the intellectual approaches to information retrieval inherited from the past, one may ask whether a preferred way exists for insuring progress in the information field. For obvious reasons, a variety of responses are given, one of which is primarily based on concern for the effective utilization and management of existing services. This approach disregards possible shortcomings in the intellectual environment and attempts to match the capabilities of the existing systems with the needs of the user populations.

More specifically, the assumption is that the sophisticated existing on-line systems are not fully exploited because of management problems, inadequate marketing efforts, and lack of user guidelines and training. [25,26] This leads to a variety of efforts in the areas of management and use of information services, and to plans for the implementation of more comprehensive automated information systems.

Under the general area of management one might cite studies aiming at the coordination of a variety of information services in diverse subject areas, or at the establishment of appropriate international information networks. The acceptance and utilization of these networks would then depend both on the ease with which access can be obtained, and on appropriate motivations among the user population. User training and education projects are thus stressed, as well as survey efforts designed to ascertain user preferences and needs. At the same time, an attempt is made to improve the coverage of existing systems by extending the scope of the service to additional subject areas, by storing factual data in addition to bibliographic information, and by encouraging automated publication efforts designed to make available documents and citations in machine-readable form.

Obviously, both the management and the use of the existing operations can be greatly improved. Anyone familiar with the administration of information services — particularly on the international level — knows that the current plans can stand a good deal of serious attention. However, this type of effort cannot alleviate the fundamental mismatch previously pointed out between the technology and the intellectual environment in information processing. Some suggestions for work in this direction are made in the remainder of this section.

B) Content Analysis and Indexing

Over the last ten to fifteen years, a great many automatic indexing systems have been developed which are capable of taking a document text or text excerpt as input, while generating as output an acceptable content representation in document vector form by fully automatic means. When such automatically generated document vectors are incorporated into a retrieval system it is found that the more sophisticated automatic indexing methods produce retrieval results that are fully equivalent with those obtainable by manual methods. [27,28]

A detailed consideration of automatic indexing is beyond the scope of this study. It is, however, important to determine where progress has occurred and why some of the early pessimism has turned out not to be justified. The first important fact about automatic indexing is that the number of content indicators assigned to a given information item is normally much larger than in the manual case — typically between 50 and 100 terms might be assigned to a given item. In these circumstances, the "depth" of indexing is much greater than in conventional situations and a large number of different possibilities exists for accessing each particular document. Furthermore, when 100 terms are present in a document vector, the accuracy of each individual term is not so crucial.

One immediate effect of the large number of terms included in the document vectors is the ease with which it becomes possible to generate similarity measures between pairs of documents (or between a document and a query). Indeed, each document (or query) may be considered to be a vector in t-space, where t is the number of distinct terms in the system, and the similarity between two vectors is then determined simply by a global comparison between the corresponding term sets. A typical similarity measure might be the cosine of the angle between two vectors — when the term assignments are identical, the cosine measure produces a similarity equal to 1.

It should be pointed out that the computation of similarities between documents implicitly defines a "document space" where items exhibiting large similarities appear close to each other, while those with smaller similarities are further removed. There is no need in such a system to consider the "distance" between individual topics, or to build a topic map including individual term relations. The main requirement is that the distances, or similarities, calculated between individual vectors correspond approximately to the similarities between the respective term sets. For the similarity measures described in the literature this is generally the case, and Mooers' three-vector example for which the triangle inequality does not hold is a pathological case that does not produce difficulties in practice.

A question arises concerning the choice of the individual terms to be incorporated into the document vectors. A variety of theories are described in the literature that can be used to choose good term sets for given subject areas, and to assign to each term included in a given document vector a weight reflecting its presumed importance. One such theory computes a term discrimination value for each term as a function of the separation which can be induced between individual documents when the term is assigned to the items of a collection. Specifically, a good discriminator is one which distinguishes the documents from each other by decreasing the average document similarity when the term is assigned to the collection items. The reverse is true for a poor discriminator which renders the documents more similar. [29,30]

The discrimination value of a term is a collection dependent measure. This must be combined with a measure of the importance of a term in each individual document to produce a weight assignable to a given term in a given document. A typical term weight might be the frequency of occurrence of a term in a document multiplied by its discrimination value. The computation of individual term values transforms the binary document vectors (where weights are restricted to 0 and 1) into numeric vectors where terms are weighted in presumed order of importance.

The assignment of values to terms extracted from individual documents is only partially responsive to Bar Hillel's criticism concerning the inadequacy of simple keyword indexing and automatic text searching. While each specific term assignment may be less crucial when a great variety of terms is available for indexing purposes, it is also true that the vast majority of the terms obtained in this fashion might exhibit very poor discrimination values. Such terms might then be worthless for content identification. Various methods can be used for the construction of improved indexing vocabularies. One such relates the discrimination value of a term to its document frequency — the frequency with which the term is assigned to the documents of a collection:

a) Terms with high document frequencies (that is, those assigned to a reasonably large number of documents in a collection) are generally the worst discriminators.

b) Terms with very low document frequencies assigned to very few documents in a collection are poor discriminators; they hardly affect the average similarity between documents when they are assigned to a collection.

c) The best discriminators are those whose document frequency is neither too high nor too low, with document frequencies between n/100 and n/10 for n documents in the collection.

A strategy for the generation of good indexing vocabularies is then one which uses the good discriminators directly for content representation without further transformation, while at the same time improving the remaining vocabulary by suitable alterations: [29-31]

a) Terms whose document frequency lies between n/100 and n/10 should be used for indexing purposes directly without any transformation; these terms include the vast majority of the good discriminators.

b) Terms whose document frequency is too high — above n/10 — comprise the worst discriminators. These terms are too general in nature, or too broad, to permit proper discrimination among the documents; hence their use produces an unacceptable precision loss (it leads to the retrieval of too many items that are extraneous). These terms should be transformed into lower frequency terms. The most straightforward way of reducing the assignment frequency of the terms is to combine them into <u>indexing phrases</u>; obviously each phrase will exhibit a lower assignment frequency than the corresponding high-frequency components.

c) Terms whose document frequency is too low — below n/100 — are so rare and specific that they cannot retrieve an acceptable proportion of the documents relevant to a given query; hence their use depresses the recall performance. These terms should be transformed into higher frequency terms. The most immediate method for increasing individual term frequencies is to assemble a number of low frequency terms into term groups and to assign such term groups for content identification instead of the individual terms. This forms a <u>thesaurus of terms</u>. Clearly the assignment frequency of a thesaurus class will exceed that of the individual components included in the class. Thesauruses are widely used to enhance the recall performance of retrieval systems (by increasing the proportion of relevant items that can be retrieved).

The foregoing theories can be used for indexing purposes by automatically assigning to the documents of a collection content indicators consisting of single terms, term phrases, and thesaurus categories in such a way that each indicator contributes maximally toward the identification of the individual documents. Since the assignment of terms to information requests is identical to that used for the documents, a retrieval system ensues with considerable discriminatory powers.

C) File Organization, Search and Retrieval

Following the generation of the vectors representing the documents of a given collection as well as the incoming user queries, the search process will generate for each query-document pair a corresponding similarity coefficient. Documents can then be ranked at the output in decreasing order of their similarity with the query, in such a way that the items located closest to the individual queries are retrieved first. The problems created by the use of Boolean connectives can be avoided entirely because retrieval may be based on a global term vector match, rather than on Boolean relations between terms. Furthermore, the size of the output may be controlled by using the ranking feature to retrieve as many items as desired.

It remains to describe the file organization and a specific search strategy that might actually be implemented. In practice, only two file organization schemes are known: the direct system in which the file is organized in arbitrary document order, and the inverted system in which relevant document identifications are listed under each allowable term. A direct system is useless for large files since the search time will be excessive when every document must be compared with the query. An inverted system, on the other hand, presupposes a static indexing situation because the information attached to individual documents is scattered throughout the file, and document vectors are hard to manipulate. The inverted system also makes it necessary to build and to maintain the inverted index.

In an automatic environment, it becomes possible to compute similarity measures between document vectors. In these circumstances, a document classification can be automatically generated by assembling into a common class all documents whose vectors are sufficiently similar. A great many automatic classification methods have been described in the literature, some of them efficient enough to be usable in practice for large collections. [32,3

File organizations based on document classes are known as clustered files. A typical search in a clustered file might be carried out in the following manner:

a) the document classification is generated for a given collection;

b) for each class, a class vector or centroid is generated from the document vectors in the corresponding cluster;

c) incoming queries are first compared with the centroids only; in a subsequent operation, all documents located in classes with high query-centroid similarities are compared with the query.

Clustered file organizations have the virtue of providing fast output (since only a small portion of the collection is actually compared with each query) without at the same time requiring the generation and maintenance of an inverted file directory. They also make it possible to carry out the search operations iteratively in such a way that initial, partial search results can be used to improve retrieval output obtained in subsequent searches.

One such iterative search process that appears to operate effectively in practice is the well-known relevance feedback method in which terms attached to previously retrieved documents are used automatically to generate improved query formulations. [32-34] A complete search can then be carried out by retrieving only a few documents initially, and using these to produce new query formulations; these new queries can in turn retrieve some more documents, leading to additional query modifications, and so on, until adequate output results are obtained. Iterative searches with automatic query reformulations require on-line retrieval environments and must be based on file organizations which make available complete document vectors for easy manipulation.

D) Dynamic File Maintenance

In conventional information retrieval environments a one-level information store is used, in the sense that all the documents are equally accessible, and the files are effectively static in nature. In practice, the file structure is often dynamic, since new items are continually added, and new interests may develop on the part of the user population requiring access to new kinds of materials. Ideally then a multi-level information store is needed in which the most important items appear "near the top" where searches are conducted most frequently and at least cost. Items which are less useful would appear in the secondary (or tertiary) stores to be searched only on special occasions at higher cost. [33,35]

Such a file organization suggests an automatic document space modification system in which important documents — for example, those that are retrieved in response to user queries and identified as relevant — are promoted little by little, while negative items — those that are never retrieved, or always judged to be nonrelevant — are successively demoted. A natural extension of such a dynamic file is an automatic system for document retirement which relegates to lower level files, items that have been sufficiently demoted. [33]

The need for document retirement arises directly from the file growth experienced in nearly all information storage environments. Unless the available resources are infinite, an indefinite growth can never be sustained, and old items must therefore give way to new incoming ones. Standard solutions to the retirement problem are the age criterion which dictates that older items with early publication dates should be removed when new items must be accommodated. Alternatively, the usage which an item receives, as determined for example by the number of citations it attracts in the literature, or by the number of times it is requested in a given time period, may determine the utility.

In a computer environment based on interactive searching and dynamic collection maintenance, file growth and retirement can be handled as a by-product of the normal search operations by adding new items to the best existing document classes, and carrying with each document special usage indicators that are automatically updated whenever a document takes part in a search process. By utilizing the dynamic file concept, the file sizes can be kept constant even when new items actually appear in the system, thus stabilizing search times and storage costs.

The dynamic file idea is difficult to implement in practice for large files in operational situations since machine-readable records are needed for all items. However, many of the other developments described in this section, including global vector matching, output ranking, automatic indexing, clustered file organization, and automatic query modification based on relevance feedback could have been implemented long ago. Many of the early objections by Bar Hillel and others would thereby have been answered, although, when all is said and done, no system — no matter how sophisticated — will ever produce perfect results under all circumstances.

References

[1] J.P.H. Batteké, D.M. Heaps, and M.A. Mercier, Canadian Water Resources Information: A Network Approach, Information Storage and Retrieval, Vol. 10, Nos. 3-4, March - April 1974, p. 85-100.

[2] Y. Bar Hillel, Is Information Retrieval Approaching a Crisis, in Language and Information, Addison Wesley Publishing Co., Reading, Mass., 1974, Chapter 20.

[3] R.A. Fairthorne, Towards Information Retrieval, Butterworths, London, 1961, Chapter 13.

[4] M. Schelle, Punch-card Methods in Research and Documentation, Interscience Publishers, New York, 1961.

[5] Y. Bar Hillel, A Logician's Reaction to Recent Theorizing on Information Search Systems, in Language and Information, Addison Wesley Publishing Co., Reading, Mass., 1964, Chapter 18.

[6] Y. Bar Hillel, Theoretical Aspects of the Mechanization of Literature Searching, in Digitale Informationswandler, W. Hoffman, editor, Vieweg and Sons, Braunschweig, 1962, p. 406-443.

[7] J.S. Melton, Machine Literature Searching at Western Reserve University, in Information Retrieval Today, Center for Continuation Study, University of Minnesota, Minneapolis, 1963, p. 89-97.

[8] C.N. Mooers, Retrieval by the Method of Proximity Transformations, Report ZTB-113, Zator Company, Cambridge, Mass., February 1958.

[9] H.P. Luhn, The Automatic Derivation of Information Retrieval Encodements from Machine Readable Texts, in Information Retrieval and Machine Translation, Part 2, A Kent, editor, Interscience Publishers, New York, 1963, p. 1021-1028.

[10] D.J. Hillman, Customized User Services via Interactions with Leadermart, Information Storage and Retrieval, Vol. 9, 1973, p. 587-596.

[11] R.A. Fairthorne, Towards Information Retrieval, Butterworths, London, 1961, Chapter 12.

[12] G. Salton, Dynamic Document Processing, Communications of the ACM, Vol. 15, No. 7, July 1972, p. 658-668.

[13] C.N. Mooers, Preface of "Towards Information Retrieval", R.A. Fairthorne, Butterworths, London, 1961.

References (cont)

[14] F.W. Lancaster, Evaluation of On-line Searching in Medlars (AIM-TWX) by Biomedical Practitioners, Report No. 101, University of Illinois, Graduate School of Library Science, Urbana, Illinois, February 1972.

[15] B.C. Vickery, The Structure of Information Retrieval Systems, Proc. of the International Conference on Scientific Information, Vol. 2, National Academy of Sciences, Washington, 1959, p. 1275-1289.

[16] E. Albani, S. Ceccato, and E. Maretti, Classifications, Rules and Code of an Operational Grammar for Mechanical Translation, in Information Retrieval and Machine Translation, Part 2, A. Kent, editor, Interscience Publishers, New York, 1963, p. 693-753.

[17] S.R. Ranganathan, General and Special Classifications, International Study Conference on Classification Research, Elsinore, Denmark, 1964.

[18] A. Kent, Machine Literature Searching at the Center for Documentation and Communication Research, in Information Retrieval and Machine Translation, Part 1, Interscience Publishers, New York, 1963, p. 525-544.

[19] V.P. Cherenin, G.A. Lavrenteva and N.Z. Zhidkova, An Experimental Information Language for Mechanized Searching of Scientific-Technical Literature, in Information Retrieval and Machine Translation, Part 1, A. Kent, editor, Interscience Publishers, New York, 1963, p. 389-428.

[20] V. Rosenberg, The Scientific Premises of Information Science, Journal of the American Society for Information Science, Vol. 25, No. 4, July - August 1974, p. 263-269.

[21] J.C.R. Licklider, The On-line Intellectual Community, in Second National Symposium on Engineering Information, Engineers Joint Council, New York, 1965, p. 29-36.

[22] C.F.J. Overhage and J.F. Reintjes, Project Intrex: A General Review, Information Storage and Retrieval, Vol. 10, No. 5-6, May - June 1974, p. 157-188.

[23] B. Lawrence, B.H. Weil, and M.H. Graham, Making On-Line Search Available in an Industrial Research Environment, Journal of the ASIS, Vol. 25, No. 6, November - December 1974, p. 364-369.

[24] M.E. Williams, Use of Machine-Readable Data Bases, in Annual Review of Information Science and Technology, Vol. 9, C.A. Cuadra and A.W. Luke, editors, American Society for Information Science, 1975, p. 221-284.

References (cont)

[25] NSF Floats New Approach to IR Support, Information Retrieval and Library Automation, Vol. 10, No. 5, October 1974, p. 1-3.

[26] Denver Research Institute, The Users and Uses of Scientific Information — Reports and Bibliographies, Science Associates International, Vol. 4, No. 1, 1975.

[27] G. Salton, A New Comparison Between Conventional Indexing (Medlars) and Automatic Text Processing (SMART), Journal of the ASIS, Vol. 23, No. 2, March - April 1972, p. 75-84.

[28] G. Salton, Recent Studies in Automatic Text Analysis and Document Retrieval, Journal of the ACM, Vol. 20, No. 2, April 1973, p. 258-278.

[29] G. Salton, C.S. Yang, C.T. Yu, Contribution to the Theory of Indexing, Information Processing 74, North Holland Publishing Co., Amsterdam, 1974, p. 584-590.

[30] G. Salton, A Theory of Indexing, Regional Conference Series in Applied Mathematics, No. 18, Society for Industrial and Applied Mathematics, Philadelphia, 1975.

[31] G. Salton, C.S. Yang, and C.T. Yu, A Theory of Term Importance in Automatic Indexing, Journal of the ASIS, Vol. 26, No. 1, January - February 1975, p. 33-44.

[32] G. Salton, Automatic Information Organization and Retrieval, McGraw Hill Book Co., New York, 1968.

[33] G. Salton, Dynamic Information and Library Processing, Prentice-Hall Inc., Englewood Cliffs, N.J., 1975.

[34] G. Salton, The Performance of Interactive Information Retrieval, Information Processing Letters, Vol. 1, No. 1, February 1971, p. 35-41.

[35] G. Salton, Dynamic Document Processing, Communications of the ACM, Vol. 15, No. 7, July 1972, p. 658-668.