INFORMATION RETRIEVAL; STATE OF THE ART

SUMMARY

Certain aspects of science communication are of especial importance to information retrieval. The exponential growth of science raises many questions related to increasing specialization. "Quality identification" is suggested as a critical issue. All approaches to information retrieval share a common set of basic problem areas and solutions. Semantics and redundancy are key conceptual issues and give rise to difficulties more likely to be overcome by meticulous thesaurus compilation than by any sudden insight or "break-through." The effectiveness of present retrieval capabilities is largely unknown, though certain recent studies are illuminating. Presentation and display to the user are suggested as important approaches to problems of information digestibility.

Information retrieval embraces only one aspect of a broad class of science communication problems. A proper perspective for assessing the state of the information retrieval art can best be achieved through considering first the broader problem context. Certain elements of that context will be discussed before the subject of information retrieval itself.

Important Aspects of Science Communication

Who Pushed the Panic Button?

The obvious fact that the world's store of scientific information is increasing at an exponential rate has apparently enjoyed independent discovery by scores of science information writers during the past few years. On a suitably scaled graph, any rising exponential curve tends to produce an hypnotic effect of impending crisis. Now an atmosphere of alarm, since it is conducive to action, is no doubt beneficial, but only provided such action is properly directed. Increasing volume of information is not in itself the real problem. It can be inferred in fact that in the context of increasing world scientific population, the information increase rate is not at all out of balance. The number of published papers, the number of journals, the number of scientists, and expenditures for research, all have been increasing at an annual rate of 5-7% for the past 250 years. Since the earth's population increases at only 1.6% annually, indiscriminate extrapolation of the exponential growth of science clearly is absurd. Limiting factors will of course set in. Prof. Bar-Hillel has challenged the view that anything at all is worsening (so far as basic long term trends are concerned), and rightly insists that the prophets of calamity must produce more convincing evidence than that which has yet come to light. Since on a scientist per capita basis the quantity of scientific information remains about constant, the case for pushing a panic button is at best obscure. The real issue must be identified as a changing partition of an increasing supply of human knowledge among the increasing number of scientists who generate and assimilate that knowledge. More simply, a scientist cannot now realistically expect to keep on top of as large a portion of any particular field or discipline (such as "physics" or "logic") as he could some years ago. The problem must be considered in terms of "backlog" as well as "current production."

Many profound questions, presently unanswerable, should be diligently explored, not in an atmosphere of crisis but in recognition of their long term fundamental importance. Should scientific journals be organized to meet increasingly more specialized needs? Does the manner of journal organization itself influence the course of scientific research? Can excessive specialization inhibit scientific creativity? Should steps be taken to "purge" the world's scientific literature, or at least to separate the important from the unimportant? Are more surveys, review papers, and monographs needed? How should education systems and information retrieval systems be kept flexibly responsive to the increasing segmentation of scientific knowledge?

These and other issues must be identified so that guidelines to future research may be developed but further pursuit of the matter is inappropriate here. So far as information retrieval itself is concerned, it will be assumed from this point on that things are bad enough and ought to be improved, regardless of the rate at which they may or may not be getting worse. The conceptual, rather than equipment, aspects of the subject will be emphasized.

Mooers' Law

Mooers has postulated that "an information system will tend not to be used whenever it is more painful and troublesome for a customer to have information than for him not to have it." By way of explanation, he observes that the penalties for not being diligent in library research are minor if they exist at all. Clearly a greater amount of recognizable and rewarding (even though duplicative) work can be accomplished if time is not spent in the library. Effective and efficient information retrieval systems have been installed and then removed for lack of use; the customer's ability to retrieve information has been stripped of his motivation. It is suggested in effect that the diligent finding and
use of information should be rewarded and failure to do so punished; in these circumstances we can expect better retrieval systems.

Though Mooers may have overstated the case, his reservations are provocative. If this state of affairs indeed exists, and certainly it must to some extent, then the question of remedial action is difficult and important. In addition to the concepts of "reward" and "punishment", two additional directions for improving "motivation" seem to be promising.

First, more attention ought to be paid to active dissemination of scientific information rather than passive interment in libraries wherein the resurrection of information requires initiative and ingenuity on the part of the customer. Perhaps the scientific community is in need of something akin to "screening panels" whose job is to direct newly acquired information to the appropriate potentially interested users.

Secondly it is suggested that beyond, and independently of "information retrieval" per se, the proper presentation and display of information to the customer may play an important role in the palatability and digestibility of the retrieved information. The extension of present information retrieval boundaries in this direction will be discussed later in the framework of research trends and goals.

Birth Control of Technical Reports

Scientists are painfully aware that much, if not most, scientific literature is either so repetitive, verbose, or of such low quality, that it ought not to have been written in the first place. In the opening address at the 1958 International Conference on Scientific Information, Sir Lindor Brown made a sparkling appeal for the exercise of restraint on the part of both authors and publishers. In professional journals, quality of information is at least partly controlled through editorial sanction, but the now comparable volume of unpublished technical reports is totally immune to such controlling influence. Dwight Gray pinpoints the lack of bibliographic control as an especially unfortunate aspect of technical reports.

It is doubtful that any system of "information birth control" can be made practical, particularly one that appeals to voluntary self restraint; involuntary control however may be a cure worse than the disease. "Quality Control" applied more liberally may be a part of the answer, but any connotation of "control" raises the question of information suppression without competent and diligent exercise of the judgmental function. "Quality Identification" is suggested here as a more appealing approach. Clearly in a situation in which the volume of information exceeds the individual scientist's digestive capacity, some method of assisting him to discriminate between the important and the unimportant literature can be counted among the most crucial of requirements.

The referee system for professional journals imposes a modest degree of quality control on scientific information. It is at least thinkable that recognized leaders in the various technical fields could be imposed upon to implement a "quality identification" system on a sizable scale. However, a more practical technique immediately at hand may lie in the idea of "citations indexing." It is not unreasonable to measure the importance of a published article or technical report by the frequency of subsequent citations to that article or report (possibly taking into account who did the citing and why.) This suggestion was ably described by Garfield in an article published six years ago. Experimental studies of citation patterns in physics literature has been carried out by Kessler.

The idea of automatic abstracting has enjoyed considerable publicity and popularity during recent years as an approach to information compression. Techniques for so doing have somehow managed to come into being without the benefit of any adequate description of requirements. The situation can be rather easily understood by first considering one plausible elementary rule for "automatic abstracting" (or rather "automatic extracting"): "Select title and first sentence of each paragraph." This rule is obviously mechanizable. Furthermore, at the time of this writing, no other rules have been demonstrated on the basis of which a convincing better abstract can be produced. The problem lies in the fact that adequate measures for the "quality" of an abstract (human or machine) have never been developed. Ideally, of course, one would like to cut the length of an article drastically without appreciable loss of information, but unfortunately it cannot yet be demonstrated that the percentage of information loss is significantly smaller than the percentage of text reduction. The question of a need for text reduction even at the expense of information loss is nonetheless open.

An interesting and constructive approach to "volume reduction" is contained in a recent Case Institute report, a part of which concerns the effect of condensation on reader comprehension of journal articles. If indeed one can somehow single out "measurable" factors in the area of reader comprehension, then certain experimental observations of this kind are crucial to questions of redundancy elimination and abstracting.

"Communication Habits of Scientists"

The scientific community itself has been the object of attention of a number of "opinion surveys" and scientific communication "behavior studies." These studies represent a commendable attempt to describe characteristics and attributes of our present system of scientific communication. A comprehensive review of several dozen such studies has been reported by Menzel. The stated purpose of this review is to display the variety of research that has been done on the flow of information among scientists; it does not attempt,
The total amount of information of this kind presently available tends to cause mild indignation when one attempts to assess the relevance of the whole matter to the implementation of practical improvements in scientific communication. In their present state, however, these "behavior" studies are intended only to be descriptive rather than diagnostic. It is premature to attempt to translate a description of present behavior into future requirements; to discern any relationship between the two requires considerable further study.

Certain fragmentary results of a rather general nature are worth noting, however, since they have an interesting bearing upon some of the questions raised here. Let us consider a few excerpts from the Menzel review.

"The amount of time chemists devote to reading on the job was found directly related to the ease of their access to scientific literature. Reading time was directly related to the availability of journals at chemists' desks, to the location of library facilities in the chemists' building, and to the existence of company library facilities."

"one of the greatest stimulants to the use of information is familiarity with its source."

"Sometimes pieces of work which have been ignored by the scientific community prove to be highly significant when someone finally stumbles upon them in the back volumes. It is suggested tentatively that it is often necessary to publicize information repeatedly, lest it fail to enter the stream of communications which will lead to its ultimate user. From the point of view of the consumer of the information, it seems sometimes necessary to be exposed to the information repeatedly before it will make an impact."

"On the basis of the information on chemists' reading time and on the number of articles abstracted in Chemical Abstracts in a given year, it was concluded that only one half of one percent of the articles published in chemistry are read by any one chemist."

"a large portion of articles read and considered useful have been met with by chance;"

"there is thus a good deal of circumstantial evidence for the hypothesis that the literature is used very much more for news than for reference."

The foregoing conclusions are provocative and clearly have some bearing on the issue of "literature search" versus "specific-information retrieval," as well as on the matter of Mooers' Law. It should be realized that the results as they stand shed no light on cause-effect relationships. It is not clear whether the use of current literature for news rather than reference reflects a requirement or whether it simply reflects the fact that current literature information systems are inadequate for reference purposes, and hence can be used only for news.

This identification of problem areas at least suggests action in two directions. First since the stimulant effect of scientific publication is known to be important, steps to enhance such stimulant capability can be taken. If this idea were coupled with "quality identification" we might reasonably equate "quality" with "stimulation ability." Once stimulating material has been identified, provision could then be made for wider circulation within the scientific community. The conclusion that current literature is not really used very much for reference, if correct, is dramatic in its implication of a requirement for a profound look at and perhaps overhaul of our literature reference systems.

The spectrum of behavior studies reviewed by Menzel does not apparently include any reports of active controlled experiments which exert a perturbing influence on the scientific community. Experiments with a fully controlled model system have been suggested by Kessler,7,10 and seem worthy of serious consideration.

The Conceptual Nature of Information Retrieval

All approaches to information retrieval are reducible to a common set of elements, both from the point of view of problems and solutions. The objective of any information retrieval system is to permit the originator of information to communicate with an unknown user at some unknown future time. We may presume that the user or customer is faced with a large volume of information, i.e., a library which by many orders of magnitude is too large to permit an exhaustive direct examination. Search to meet whatever requirements for information he might have. Thus communication from originator to "user" must take place via some abbreviated "representation" of the contents of the library. This representation may take many forms, for example: classification systems, alphabetic subject indexes, coordinate systems, facet classification systems, full text, and others yet to be invented. They all incorporate certain common problems, and these problems have a common origin.

When any document is indexed, cataloged, classified, or otherwise tagged, this process constitutes an attempt to represent "information content" of the article or document in a highly abbreviated form. This "representation" is intended to serve the purpose of information retrieval but contains only a minute fraction of the information contained in the document. Now it is not usually supposed that the "representation" should reflect the total "information content," yet at the same time there is no solid theoretical reason that the purpose of information retrieval can be fully served in any other way. Fortunately, the whole matter need not rest
on theoretical proof since the objective at hand is of a purely practical nature.

As a practical matter, the encoding of "information content" or "meaning" of a document in any highly abbreviated representation is an intuitive process not amenable to formal description. Experiments have indicated that the reproducibility of the intuitive judgment of indexers or catalogers is relatively low. A subject heading, index term, or descriptor, has a "meaning" which is a function of the viewpoint of the observer or equivalently a function of the context in which that heading, term, or descriptor is utilized.

The effectiveness of an information retrieval system depends on the success with which the indexer and user can pursue the following strategy. In creating an "abbreviated representation" of an article being indexed, an attempt is made by the indexer to foresee essentially all ways in which some unknown user might wish to recover the information. A rather high degree of redundancy is a result of a deliberate attempt on the part of the indexer to predict the customer's viewpoint. Similarly the search is conducted on a redundant basis since in effect the user hunts in all likely (and in some not-so-likely) places. Both indexer and user must therefore play a "guessing game" to ascertain the viewpoint of the other. This game results in redundancy, but not so much so that large quantities of retrieved irrelevant information thwart the purpose of the search. The foregoing description applies to a successful information retrieval system. The fact is, it is possible to find many systems which seem to have no elements of such guessing or redundancy and which, at the same time, are not very successful.

Consider for the moment a retrieval system in which the library "representation" is based solely on key words or descriptors. For this situation, the "guessing game", played by both the indexer and searcher, can be aided by means of a thesaurus. For our purpose here, we define a thesaurus as a collection of groups of words wherein the various members of a group tend to mean about the same thing for purposes of information retrieval. These word groups thus serve as a reminder list to assist the indexer and searcher in constructing how the same idea might be expressed in many ways. (Experimental use of a thesaurus is described for full text search in references 11, 12.) In one form or another, something akin to a thesaurus must be implicitly present in any good retrieval system. The "see also" portion of an alphabetic subject index or of a classification system can be looked upon as serving the purpose of a thesaurus.

The total amount of desirable redundancy in any retrieval system, and the detailed manner in which redundancy should be distributed between indexing and searching, depends upon economics and on the question of irrelevant retrieval. For a system in which the encoding process is highly efficient, redundancy may be cheaper to come by as a part of the input rather than as part of the output. In other systems, however, such as those in which the search workload is light, it may be that redundancy can be purchased more cheaply in the searching process.

An important problem area, more or less distinct from questions of meaning, has not been covered in the foregoing discussion. This area concerns syntactic relationships among the various index terms that might be assigned to a given document. The effect of failure to utilize such syntactic relationships is to increase the retrieval of irrelevant material but with no loss of that which is relevant. To illustrate, if the phrase "the effect of radiation on mutation" is an appropriate description of the content of an article, then the independent assignment of descriptors corresponding to "radiation" and "mutation" would not by itself preserve the syntactic relationships between these two terms implied by the original phrase. Irrelevant retrieval of documents in which the descriptors "mutation" and "radiation" were simply included among a number of other descriptors in such a way that those two were not related to one another, could clearly result. The degree to which this whole question is of importance in terms of the amount of irrelevant retrieval that might be caused by ignoring syntax has not been experimentally established. In principle though the problem is of considerable importance and a number of approaches to its solution should be mentioned.

First of all, a considerable amount of work has been in progress for some years at the University of Pennsylvania under Z. Harris. 13 The objectives of this effort are broader and more fundamental than the question of preserving syntactic relationships among index terms, but the subject is nonetheless of considerable relevance. Eugene Wall14,15 has presented a lucid categorization of information retrieval problems in terms of viewpoint, generics, semantics, and syntax. He describes the use of "role indicators", as a solution to problems of syntax, and claims that effective and successful technical information systems which employ only twelve such indicators have been in operation.

So far as full text searching is concerned, an especially simple substitute for syntax, namely, the "proximity" of terms has been suggested by the author. 12 This approach depends on the fact that two terms which appear in the text of an article in reasonable proximity to one another (i.e., within a few sentences) have a high probability of being syntactically related.

"Effectiveness" and "economy" can be identified as the two fundamental objectives of information retrieval research. "Effectiveness" has to do with how well a system works, in terms of both the percentage of relevant material retrieved and the accompanying amount of irrelevant
machine-readable coded representation of a document. This merger imposes awkward restraints (accompanied possibly by titles or brief merged records) of the document itself. Retrieval of the document should be assessed in the light of its objectives, and the division of objectives into the two categories of effectiveness and economy usually provides a perspective not otherwise apparent.

The issue of mechanization is, of course, tied to economy. It is obvious, though sometimes overlooked, that computing machines cannot in principle improve the "effectiveness" of anything at all over what can be done with people, unless response times or environmental factors are of importance. The basic question is one of economy alone. Speed itself may be of some relevance but, in principle at least, one can duplicate the speed of any automatic process with enough people working in parallel. Admittedly this argument is oversimplified (though the remark on response times and environment is a strong hunch), but justified as a counteraction to the unfortunately prevalent belief that mechanization of a process that doesn't work to begin with will improve matters.

Apart from its intimate relationship to economy, one other especially important factor in mechanized information retrieval should be recognized. The machine handling of the physical contents of a library involves problems totally different from the machine handling of a representation which permits that library to be searched. Present general purpose computing equipment, provided it is used with some ingenuity, is not badly suited to machine search of representations, i.e., index and catalog type information. So far as handling the total information content of a library is concerned, the required storage jumps several orders of magnitude, and existing computers are largely inappropriate. The output of computers which handle and search only an index-representation is necessarily just a bibliographic listing of responsive documents (accompanied possibly by titles or brief abstracts). Retrieval of the document itself (from shelves, filing cabinets, or microfilm storage) must then follow as a second stage process. Special purpose retrieval machinery can be relatively efficient for this second stage, though generally speaking well designed manual filing systems offer serious economic competition.

Some equipment designs have been based on a merged record (usually recorded on film) of the machine-readable coded representation of a document with a non-machine-readable full text document image. This merger imposes awkward machine design problems from the beginning and must depend for its final justification on overall economy of the resulting system. Up to the present time large scale systems based on the merger philosophy have not been conspicuously noted for economy and simplicity (to risk a serious understatement). (That fact was of course known from the design stage on--what's worse though, at least some of these systems after installation have been victims of a rather severe onset of Moores' Law). The question of mechanization in libraries is full of economic pitfalls, but not to the extent of preventing soundly designed systems to be implemented with present hardware. Since this paper is intended to focus on conceptual problems, the matter of equipment will not be further pursued.

A considerable amount of reported work in the information retrieval field is indirectly addressed to the question of economy rather than effectiveness though not in an obvious sense and not necessarily in connection with mechanization. A number of ingenious logical-mathematical models of information retrieval systems have been devised. Many of these papers have interesting implications on questions of file organization and search strategies, (e.g. Estrin18, Moore19) and possibly on the design of future machines; their promise for leading to new insights with respect to the more basic semantic problems that lie at the core of "retrieval effectiveness" is less clear.

It is at least plausible, however, that some of the mathematical models might eventually lead to fruitful results in a more fundamental sense. Moores20 has developed a model in which certain mathematical properties of different types of retrieval systems can be compared. It would seem useful to extend a model of this kind to include formal representation of redundancy, and then to investigate the relationship between retrieval effectiveness and redundancy. In the terminology of Moores' model, the transformation relationships between the space of all retrieval prescriptions and the space of all document subsets should be formulatable in such a way that the effect of redundancy is clearly brought out. It would be overly bold to predict at this point whether such an approach would yield any new fundamental insights, better file organization and storage strategies, both, or neither.

A key element in understanding the "nature of information retrieval" must certainly be the types of questions posed by users of information systems. It is in fact far more reasonable to design library representations on the basis of the way in which the users tend to organize the subject matter rather than the way in which indexers imagine that it ought to be organized, yet it seems that this procedure is seldom followed. This approach is embodied in a study carried out by Hermer21 on the information system of the U.S. Atomic Energy Commission. (Several interesting results of immediate practical implications, so far as machine design is concerned, were obtained. Ninety percent of the questions involved three or fewer distinct
concepts and of all of the multiple concept questions, ninety-eight percent involved logical products rather than logical sums or differences.) In connection with the Herner study, it would be interesting to know the extent to which the user's questions were predicated on some prior notion of how the AEC library was organized; that is to say, did their questions really reflect what they wanted or were they conditioned by what they thought the library could provide?

How Effective are Present Information Retrieval Systems?

It is a curious fact that the above question is essentially unanswerable in terms of any objective "measurables." For any given information retrieval system, those concerned with it are generally not at a loss for an opinion as to how well it works, but such opinions are seldom accompanied by evidence. Some recent experimental research has been seriously addressed to this point, and insight is being acquired. Large scale experiments which attempt to measure retrieval effectiveness and at the same time compare various factors crucial to such effectiveness, are presently under way (under the direction of Cleverdon) at the College of Aeronautics, Cranfield, England. Preliminary results have been reported.10 Within the framework of a small scale experimental system, considerable attention is given to the question of defining measures for retrieval effectiveness in the author's investigation of text searching.11 The results of several retrieval methods are compared with "direct examination" of the whole document collection.

The emerging results reported in these two recent studies seem to be taking on a rather interesting pattern. The AGIBM Cranfield project had for its objective the measurement of three variables: the indexing system (four specific systems were compared), the time spent indexing, and the experience or qualifications of the indexer. Preliminary results presented by Cleverdon in a recent seminar17 indicate that it doesn't much matter which indexing system is used, how much time is taken in the indexing process, or whether the indexer did or didn't have a lot of experience. Nothing seemed to depend critically on the searcher either. Similar "invariances" were encountered in the reported text searching experiments.11

On the whole retrieval effectiveness in these experimental systems was relatively low (Cleverdon reports recoveries in the neighborhood of eighty percent but it should be noted that these figures refer to "source" documents--i.e., those which directly inspire the retrieval question. Retrieval percentages for non-source documents have not yet been reported for that project. The same distinction between source and non-source documents was used by the author,11 and it was found that recovery of source documents was about twice as effective as that for non-source documents.) In general these experiments seem to indicate mediocrity of retrieval effectiveness and insensitivity to parameters that might reasonably be supposed important. Though further analysis is necessary before firm conclusions can be drawn, certain hypotheses to explain the observed mediocrity and insensitivity can be suggested. If we suppose that the thesaurus used by the author and a "see also" structure of the indexing systems tested by Cleverdon were about equally primitive (insofar as covering any substantial range of language redundancy is concerned) then one would have expected the results to come out about as they did. It has already been found that a substantial portion of the text searching ineffectiveness could reasonably be attributed to a deficient thesaurus.12

An interesting sidelight brought out by Cleverdon in the NSF Seminar is that there presently exists no book or well organized doctrine on how to compile subject heading lists nor does there exist a systematic body of opinion on the application of "see also" references. Cleverdon observes that the latter particularly tend to be applied in a haphazard manner. With inadequate thesaurus groups, cross references, and haphazard "see also" application, it should come as no surprise that the indexer and user find difficulty in communicating via an information retrieval system.

A few remarks can be made on the state of automatic indexing. It has been pointed out that the language redundancy problems associated with retrieval effectiveness must be approached through thesaurus-like compilation techniques. On the basis of all present evidence, there is no particular reason to believe that the process is any more difficult for automatic indexing than it is for human indexing. The more subtle aspects of semantics and redundancy in the system based on automatic indexing must in that case be thrown into the search process, but with no obvious disadvantages. The real question is not so much whether automatic indexing can be made to work (though to be sure the matter must still be left open) but whether it can be made economical. Generally speaking, with present computing equipment it cannot, except for certain special applications where input costs can be amortized in other ways. Engineering achievements in the area of direct printed page input and higher speed memory readout, may hold the final answer to this question.

Trends and Goals

The nature of the basic problems of information retrieval is such that no sudden conceptual breakthrough seems likely. Current inventories of mathematical theories and techniques are applicable to information systems only to a limited extent. The tasks that clearly lie ahead must include large scale language studies and laborious experimental investigations. The past several years have seen increasing awareness of the significance of some kind of thersaurus approach to problems of multiple meaning, viewpoint, generics, and redundancy. This approach is not limited to natural
language search techniques or to key-word descriptors, but finds its counterpart in the "see also" cross references of all types of subject headings and classification systems. These compilation efforts which in effect attempt to build "structures of relatedness" in indexing and retrieval systems should be closely tied to studies of the types and forms of questions which users ask of the system. Studies of users' questions should be patterned not only after the kind that are presently asked (e.g., hemer21) but should be addressed also to the kinds of questions which users ought to ask if they had different and better information retrieval systems.

Retrieval techniques work best when they deal with a limited subject area and are tailored to the requirements of a limited group of users. Working systems of this kind should form the nucleus for experimental investigations so that deeper insights can be obtained into the question of how and why they behave as they do.

Syntactic studies should continue though it may be anticipated that their practical relevance to problems of information retrieval may not materialize to the degree that many workers presently hope. It seems to the author that problems of semantic transformations of one another and this fact alone suggests that syntacticians may be eventually disappointed in the extent to which their work finds practical application. Storage and retrieval of condensed or "kernelized" sentences13 suffers from the same illness as do other methods of automatic abstracting; at present there are no acceptable measures for the amount of information loss in the kernelization process. This entire area may well emerge as being of considerable future importance however if for no other reason than the keen interest of a considerable number of competent workers; it holds much potential for taking off in new directions.

The presently primitive state of the automatic indexing art was mentioned briefly in the last section, and it is clear that studies of this kind shall and should continue. It is in this area that advances in equipment capabilities are urgently needed. A portion of these studies should be addressed to the interplay between natural language queries and mechanized information retrieval systems. The formulation of a request in natural language for computer processing is not subject to the current limitations of high volume storage that natural language text searching involves.

The extent to which one should be sanguine about continued studies of the information gathering habits of scientists, as though they were a colony of bees or ants, is not really clear. Certainly a rash of these studies has broken out during the past four years or so and many of the results have been interesting. It is possible that these have largely run their course and that now emphasis ought to shift to experimental studies that are subject to greater control. Certainly more investigation should be made of the obviously observable parameters within the science communication sphere such as those provided by the technique of citation indexing. Other techniques exist for "hooking" documents together by purely pragmatic and intuitive estimates of relevance within the framework of some particular purpose. A set of citations as useful "relevance hooks" would be based on an assumption that co-citation implies similarity of meaning from the point of view of the author doing the citing. In an analogous sense, Fano22 suggests that the recognition of similarity of two documents in the process of request formulation (i.e., the requestor asks for a document "like" one he already has) provides relevance hooks similar to those provided by co-citation. Networks of relevance chains once assigned are susceptible to highly efficient machine manipulation and have the added virtue that the profound questions of semantics, viewpoint, ambiguity, generics, and syntax, are involved only to the extent of directly observable external effects resulting from a large collection of instances in which professional judgment is applied to the question of "relatedness."

It was mentioned earlier in connection with Mood's Law that extending the boundaries of information retrieval into the area of presentation and display may provide an approach to solving problems of information indigestion on the part of the user. Strictly speaking, an information retrieval system has served its purpose once responsive documents have been delivered to the customer, but if indeed the customer's attitude is characterized by indifference, to the extent that the retrieval system is not effectively used, then we might surmise that a highly important and critical problem area has been ignored. This area is one which begins where information retrieval ends. Let us imagine that a stack of several dozen documents of several thousand words each has been retrieved. Now the customer's original motive in requesting those documents must be examined. If he seeks specific pieces of related information, a considerable effort on his part may be required to extract what he wants. If that is indeed the case then further machine processing of such retrieved data may serve a useful purpose. It is of particular significance to note that computer handling of the total retrieved text (in contrast to the total text of the library) is reasonable in terms of storage requirements. His original requirement may of course have been of a different kind, such as acquiring general familiarity with a subject area, but we shall consider further only that situation which calls for specific small fragments of information to be brought together and "correlated" in some abstract intellectual way. This type of requirement typically occurs in a business or military intelligence application. The question now is whether or not the proper presentation, rearrangement, and display of numerous fragments of retrieved data may play a significant role in stimulating the user to...
6.1

perceive relationships that are not otherwise apparent in a more laborious process of directly examining the total contents of the retrieved material. Experiments carried out by the author and co-workers at Ramo-Wooldridge during the last year or so have demonstrated that such stimulation is indeed possible. (A report on this work is in preparation.) When combined with one of the conclusions quoted in the Menzel review, namely that the periodical literature tends to be used more for "idea stimulation" than for reference, these concepts of presentation and display take on added significance. It is suggested that future information retrieval work may exhibit an interesting trend in the direction of automatic user stimulation.

The articles cited in the following list of references themselves contain references to 160 other articles (probably some are duplicative). If succeeding generations of citations are counted, one wonders whether a closed system will be encountered (it seems likely) and whether subsets of those cited many times are in some way more "central" or "important" to information retrieval.

REFERENCES


2. The Real Problems Behind Information Retrieval and Mechanical Translation, Y.Bar-Hillel, Seminar Feb. 13, 1961, Sponsored by the Office of Naval Research


4. Opening Address, Sir Lindor Brown, ICSI* p. 3


8. An Operations Research Study of the Dissemination and Use of Recorded Information by Operations Research Group, Case Institute of Technology, December 1960 (Sponsored by National Science Foundation)


10. Background to Scientific Communication by M. M. Kessler, IRE Translations, Vol. EWS-3 No. 1 April 1960


13. Linguistic Transformations for Information Retrieval by Z. S. Harris, ICSI* p. 937


15. Summary of Area 4 Discussion, ICSI* p. 804, 805

16. ASLIB CRANFIELD RESEARCH PROJECT, Report on the First Stage of an Investigation Into the Comparative Efficiency of Indexing Systems by Cyril W. Cleverdon, Sept. 1960 (Sponsored by the National Science Foundation)

17. Seminar February 17, 1961, by Cyril W. Cleverdon, held at National Science Foundation, Washington, D.C.

18. Maze Structure and Information Retrieval, Gerald Retin, ICSI* p. 1383

19. A Mathematical Theory of Language Symbols in Retrieval, by Calvin N. Mooers, ICSI* p.1327

20. Determining Requirements for Atomic Energy Information from Reference Questions, by Saul Herner and Mary Herner, ICSI* p.151

21. Summary of Area 6 Discussion, ICSI* p.1407