JOURNAL OF RESEARCH of the National Bureau of Standards — B. Mathematical Sciences
Vol. 72B, No. 1, January–February 1968

Citation Searching and Bibliographic Coupling With Remote On-Line Computer Access*

Franz L. Alt and Russell A. Kirsch**

Institute for Basic Standards, National Bureau of Standards, Washington, D.C. 20234

(January 30, 1968)

Experiments were performed on a remote, multiple access computer to retrieve bibliographic items based on citation data. The citations were from some 25,000 physics papers. Estimates were made of the relevance of bibliographies derived from such citations. Some extrapolations are made to future systems with remote multiple access capability.

Key Words: Bibliographies, citations, computer, information retrieval, remote consoles, timesharing.

1. Introduction

The use of citation indexing for information retrieval is increasingly arousing the interest of the scientific community. It therefore seems appropriate to document some of the experience which the authors have gained with this tool.

A main purpose of the experiments reported here was to determine whether a citation index can be used effectively by remote access to a computer-stored file, and how this process compares with conventional use in a library. A computer-stored citation index has some obvious advantages; it seemed important to find out whether it involved any unforeseen difficulties or, for that matter, any unexpected benefits.

We shall first discuss citation indexes in general, their characteristics and the use that is made of them. This will enable us to appreciate the built-in advantages of a mechanized index. We shall then describe a number of experiments which we have conducted, and finally enumerate a few conclusions which can be drawn from them. These experiments had the special feature of using the computer remotely; it is this characteristic which distinguishes them from earlier trials of mechanized citation indexing [1, 2].[1]

Remote access to computers [3] is made attractive by two kinds of considerations, which in the literature are not always clearly distinguished, although they differ widely in their effects on the design and economics of the computer system. In one class of cases, long-distance access to a computer is desirable because the computer stores a unique information file (or program). The alternative of storing a copy of the file in a locally used computer may be impractical because of frequent need for updating or revision, or may be uneconomical because the file is so large and/or so infrequently used that the cost of computer storage exceeds the cost of long-distance communication. In another class of cases, the desire for man-machine interaction is the controlling consideration. To match the speed of humans and machine calls for time-sharing, usually among several dozen users (even though the machine does the lion's share of the work). It is normally not practical to assemble so many users in the machine room; but in contrast to the former use,

*Work on this project was supported by the National Science Foundation grant number GN–320.

**American Institute of Physics, 335 E. 45th Street, New York, N.Y. 10017.

[1] Figures in brackets indicate the literature references at the end of this paper.

it is often possible and desirable to share the computer's time only among users in the same city, or even in the same installation, thus almost eliminating the costs of communication. In our work with citation indexing, both considerations—access to a unique file and desirability of man-machine interaction—are present.

2. Characteristics and Use of a Citation Index

A citation index ([4], and refs. cited there) is a list of documents (e.g., scientific papers) which cite (make reference to) other documents. Its main purpose is to find citations to a given paper. The index must contain at least an identification of each citing paper and similar identifications for all cited papers associated with each citing paper. The identification should be unique and unambiguous, or nearly so. There may be additional information if desired; though if there is too much of it, the file will be less economical in serving its primary purpose. The order in which the list is arranged, and the format of its items, also have a decisive effect on the economics of the operation. We shall have more to say about this later.

Figures 1 and 2 show typical entries from two citation indexes. The former is taken from the file of the Technical Information Project (TIP) at Massachusetts Institute of Technology (MIT). The first line is the identification of a *citing* paper; in it, J001 is a symbol arbitrarily assigned to one journal (in this case, Physical Review), the remainder of the line gives the volume and (starting) page number. This identification is obviously unique, and it is unambiguous except in the rare case where two papers start on the same page. Similarly, the block of symbols at the end of the entry gives the identifications of the 38 *cited* papers. (Journals are assigned serial numbers lying between 1 and 999; references to papers not in one of these journals are omitted.) Additional information given in the index consists of title of the citing paper, author(s) and his (their) affiliation(s). There is no additional information for the cited papers. Figure 2 is from the Science Citation Index (SCI) of the Institute for Scientific Information. The identification for both citing and cited

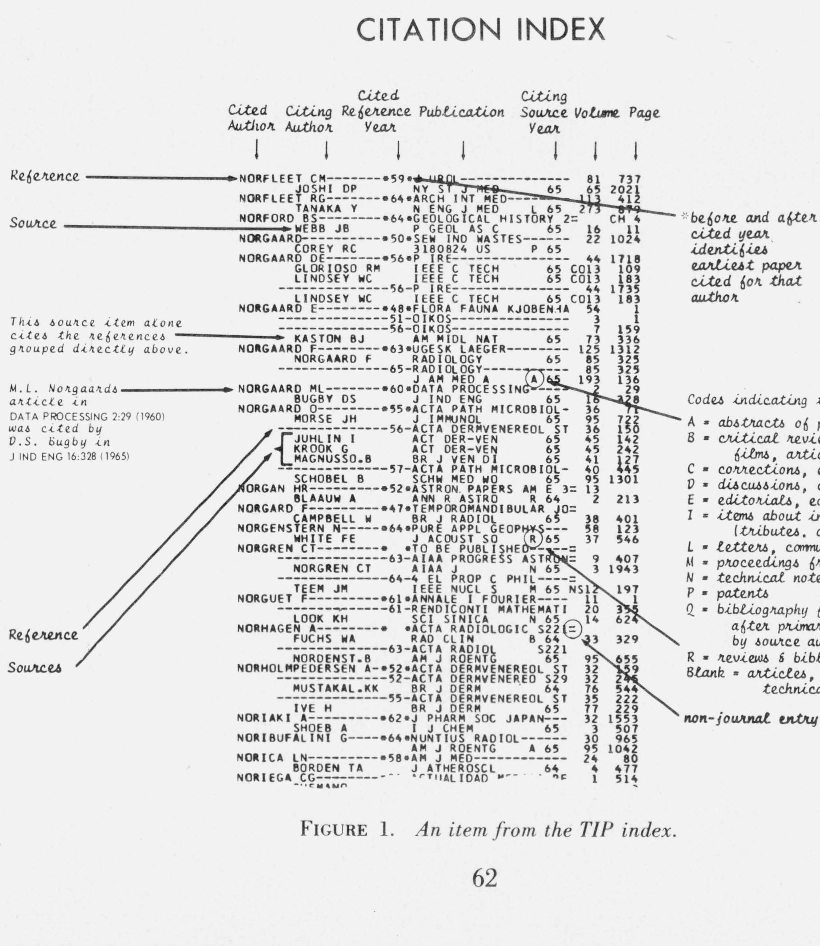

FIGURE 1. *An item from the TIP index.*

```
PHYSICAL REVIEW
VOLUME 150
J001 V150 P0429
RELATIONSHIP BETWEEN THE AUGMANTED-PLANE-WAVE AND
KORRINGA-KOHN-ROSTOKER METHODS OF BAND THEORY
  JOHNSON KEITH H.
      J001 V051 P0846    J001 V092 P0603    J001 V092 P1126
      J001 V094 P1111    J001 V105 P0108    J001 V124 P1786
      J001 V124 P1797    J001 V125 P0109    J001 V128 P0082
      J001 V131 P2529    J001 V139 P0760    J001 V145 P0599
      J002 V000 P0000    J002 V000 P0000    J002 V000 P0000
      J002 V000 P0000    J002 V000 P0000    J003 V086 P0337
      J003 V086 P0825    J019 V042 P0276    J021 V013 P0392
      J027 V000 P0000    J027 V000 P0000    J027 V000 P0000
      J027 V000 P0000    J027 V000 P0000    J384 V064 P0513
      J384 V066 P0545    J384 V072 P0345
```

FIGURE 2. *A segment of the ISI citation index.*

papers consists of abbreviated journal title, volume and page number; the first author's name and the year of publication are given as additional information for both citing and cited papers. The TIP file is stored in randomly accessible computer memory, from which figure 1 is printed out; the SCI is derived from a file on magnetic tape but is accessible to outsiders mainly in printed form. TIP is arranged by journal, volume and starting page of *citing* papers; SCI is arranged by year of citing paper,[2] and within each year alphabetically by name of first author of *cited* paper.

In TIP, each citing paper is followed by its cited papers. SCI, on the other hand, lists each cited paper, followed immediately by pertinent citing papers. (The Institute for Scientific Information also maintains files in several other arrangements.) It could be argued that the term "citation index" ought to be reserved for the latter form, because unlike any other index, it responds immediately to the basic query "find citations to a given paper." We shall, however, use the term more broadly as applying to any list (such as the TIP index) from which this basic query can be answered without unreasonable effort.

Let us examine the last point more closely. Any request for information from a file may be considered as consisting of at least two parts, of which one specifies the "search-range," i.e., the portion of the file which is to be searched, while the other gives criteria for deciding, for each item in the searched portion of the file, whether the item is or is not wanted as part of the answer to the request. In manual searching these two parts are implicit in the actions of the searcher, where he looks and what he looks for. In the TIP file, where searching is done by computer, they are formalized as two computer statements ("macro-instructions"); one begins with the word **SEARCH,** followed by a list of journal volumes to be searched; the other begins with **FIND** and specifies, e.g., some cited papers; other types of "find" specifications are possible and often useful, such as author's name, words in title, etc. The two parts of the request look as if they played entirely different roles, but logically speaking their functions are quite symmetric. Let A be the set of all papers in the search range, B the set of all those papers in the file which satisfy the "find" specifications; to satisfy the request we have to form the intersection of the sets A and B. The request "search B, find A" would have the same result as "search A, find B." For instance, a request to the TIP index might specify "search Phys. Rev. vol. 135, find citations to Phys. Rev. v. 126 p. 146," the latter being a paper by J. P. Auffray on the magnetic susceptibility of the hydrogen molecule. The result would include the item shown in figure 1, since this paper is in Phys. Rev. (J001) vol. 135, and includes among its citations the one to J001 V 126 P 146. The same request could be made to SCI in the form

[2] To the extent that the file is cumulated and printed annually. This annual cumulation coincides partly but imperfectly with an arrangement on citing year.

289-930 OL-68—5

"search citations to Auffray (Phys. Rev. Vol. 126, p. 146), find citing papers in Phys. Rev. vol. 135." Which of the two forms (or of other possible forms) is chosen will depend on the organization of the file. In TIP all *citing* papers from one journal volume are stored contiguously; therefore the search requests are composed of specific journal volumes. In SCI all *cited* papers of one author are listed in sequence. The latter arrangement is more economical when, e.g., searching for all citations to one given paper, regardless of time; the former is advantageous in requesting, e.g., all papers in the most recent literature citing one of a list of papers, or in limiting the search to journals from certain countries, or journals specializing in certain fields.

3. Types of Requests

In order to appreciate the requirements to be met by a citation index, one should first visualize the modes of consulting it and the reasons for them. For our purposes, the main distinctions will be whether the index is consulted only once for each question or periodically for the same questions; whether questions are asked singly, in batches, or in sequence; what is the time period within which the desired citations should have appeared; and what are the boundaries of the field of knowledge spanned by the citations.

One frequent reason for consulting a citation index will be a scientist's need to obtain the most recent answer to a specific question, when this answer is being revised from time to time. For instance, he may wish to get the latest value for the atomic weight of some element. He knows that this was measured some years ago, and he suspects that it may have been revised since. He enters the index with the latest publication on this subject known to him—perhaps 5 or 10 years old—in the hope that the publication of a subsequent revision would reference the previous result. In such a case the question is asked only once and is a single question (though it may be combined into a batch with other similar questions); the field of knowledge is quite narrow; the time interval of interest extends from the previous paper to the present, with emphasis on the most recent literature. If the scientist continues to work on the same element, he may consult the index periodically with the same question, in which case the time interval of interest runs only from the previous consultation to the present. An organization devoted to up-to-date knowledge of such information, say a data center, may have batches of similar questions, asked periodically, and covering a broad field of knowledge.

A very different problem is that of compiling a bibliography of a subject. Here the researcher may start with a few publications known to him, look up the references cited in them, use this composite list of references to enter the citation index and obtain a longer list of citing papers (which includes his original list), and iterate this procedure several times. At each step it will be desirable to comb out those papers not sufficiently germane to the field. In this application, there is a sequence of consultations of the citation index, each dependent on the results of the previous one and on the inspection of these results by the searcher. The time interval of interest extends over the entire period spanned by the index. There is no special premium on the most recent literature, but it is essential that the period covered be long enough; at least, say, five years, preferably longer. The field of knowledge covered by the index must be at least a little broader than that which is to be spanned by the bibliography, and it is important that this field be covered as completely as possible, including relatively obscure sources.

These two examples will suffice for our argument. There are other important applications of citation indexes—current awareness programs, patent searches, finding reviews of, or corrections to, a given paper, bibliographic coupling, etc. They merely confirm the conclusion that a citation index, to satisfy all major classes of users, must have the following properties: (a) It must cover a continuous time period of at least five years for citing papers. (b) It must be reasonably up-to-date. (c) It must cover its intended field of knowledge as completely as possible, including obscure sources in the field itself and sources in tangential fields which are apt to cite, or be cited in, papers in the principal field itself.

Point (b) requires that there be frequent additions to the index. It is likely that point (c) will require the same thing since it may be difficult to achieve the needed coverage except in successive approximations. Points (a) and (c) imply that the index has to be of large volume; a small incomplete or short-range index has disproportionately small usefulness. The size of the index makes the process of incorporating frequent additions more difficult.

The reason for desiring a coverage of five years is that experience has shown that most citations to a given paper come within five years of its appearance; thereafter the frequency of citation falls off. Thus, with a coverage of less than five contiguous years we are likely to lose some valid retrieval clues.

While neither the SCI nor the TIP index satisfy these requirements completely, both are close enough to make extrapolation of our results to a hypothetical index with completely adequate coverage permissible. The TIP index begins in 1963 for most of its journals and is thus approaching the five-year condition. Indeed, some older volumes are on magnetic tape but are not in computer memory, because of limitations of memory capacity. In SCI the lack of the years 1962 and 1963 may tend to blur the picture; the years 1961 and 1964 ff. are covered. Both indexes seem to be "reasonably" up to date; our experiments did not involve this feature. What time lag exists in SCI is perhaps the unavoidable minimum for manual updating; with TIP a speed-up in updating should be possible. Coverage of the field is distinctly better for SCI than for TIP; but it is not nearly complete for either, while on the other hand the most frequently read periodicals are covered by both.

4. Advantages of a Mechanized Index

The largest citation index in the physical sciences is the one being compiled by the Institute for Scientific Information in Philadelphia. It takes 8000 pages to cover the year 1965 alone, yet this 8000 page coverage of 1600 journals in all scientific areas does not include the titles of all journals being published. (The sizes of the journals included in these 1600, however, are such that a much higher proportion of the total number of items published is included in this coverage than the number of journal titles *per se* might imply.) Such coverage will be satisfactory for some purposes, e.g., for a current awareness program, while it would probably be inadequate, e.g., for efforts to keep track of the world literature on a special subject.

Let us, however, assume that we are satisfied with this coverage as regards subject matter, and consider the problem of coverage in time. The index consists of separate quarterly volumes and cumulative annual volumes for 1961, 1964, 1965, and 1966. A similar index is to be compiled for succeeding years. Thus it will normally be necessary to look for each desired item in all the cumulative annual indexes, and separately in each quarterly supplement which has appeared since the last annual index—a total of perhaps 4–7 places. Yet, to issue supplements quarterly is hardly enough; it be currently up-to-date the supplements would have to be issued at least monthly, if not weekly. This would increase the number of places in which one must look for each item, or else cumulative indexes have to be issued more frequently. But a cumulative index providing the subject coverage of ISI and covering five years will fill about a foot of shelf space. How frequently should this much paper be thrown out and reprinted, mailed, accessioned by librarians, and physically replaced on bookshelves?

Furthermore, consider the effort of looking up an item in a list which covers a whole bookshelf. In the compilation of a bibliography, quoted above as example, a number of items have to be looked up initially. Each may be followed by several citations which have to be manually copied, examined, their references looked up, etc. All this is made prohibitively clumsy and time-consuming by the size of the index.

One might think of working with smaller indexes by having a large number of specialized indexes, each covering a certain subfield. This could work up to a point, but as specializations are made narrower, the number of citations leading from one subfield into another increases, making it necessary to look up each item in several indexes.

All these difficulties are remedied if the index is not presented in printed form but as a computer-readable record.

First, the problem of having too many supplements or having to update the master index too frequently is easily overcome in either of two ways. A computer can search several files (master and several supplements) simultaneously and use hardly more effort than in searching a single file; and it can merge the several files and produce a new updated master file, while performing one of its regular search routines, and expend hardly any additional time or effort in the process. Thus, the master file can be kept essentially up-to-date at all times; the only information items not yet incorporated into the file are those which have been acquired since the last consultation of the file, and these will be incorporated simultaneously with the next consultation.

Second, completeness of coverage is more easily approached. It is likely that the coverage of the index will have to be gradually broadened and intensified; it would be undesirable to withhold use of the index until citations from the entire desired literature have been recorded. Thus, the gradual attainment of adequate coverage will involve repeated additions to the file. In a mechanized file these can be handled easily as above.

Third, and above all, the considerable effort of manually copying relevant references out of a printed index is avoided. The computer not only finds the desired references but provides a typewritten list of authors, titles, journals, etc. Especially in the repeated back-and-forth process involved in compiling a bibliography, this is an inestimable advantage. In addition to producing a typed list of references at each stage, the computer retains the same record in its memory. The searcher can inspect the typed copy, indicate to the computer which items should be deleted (a convenient service program needs to be available for this stage), and the computer is automatically ready for the next step. Observe that on-line interaction between the questioner and the machine is desirable in this process.

These advantages are avialable, more or less, with any computer-readable citation index, regardless of the machine which is used. In our case, further benefits accrued from the special facilities of the CTSS time-sharing system of Project MAC at MIT [5], in which TIP is embedded. For example, there are simple system instructions for creating and editing files, checking them and correcting errors, printing in different formats. A single instruction will insert a change in every place where it applies. Temporary files can be created when the user has run out of permanently assigned memory space. Other system instructions monitor the operation, tell the user at his request how much time and storage space he has used etc. Different users communicate with each other by a programming scheme called "mail box," which alerts them to messages stored in the computer by one user for another; they can also, subject to an elaborate system of file protection, gain access to each other's files for data and programs. All these features were used to good advantage in our own work. There are others which, although not used by us in the present context, could be most useful in other work in citation searching. CTSS implements programming languages like COMIT which make it easy for the user to supplement the TIP instructions by programs of his own choosing for reformatting files, obtaining statistical results regarding searches, or combining citation searching with other search methods.

It might be objected that many researchers do not have easy access to computers. This problem is not as serious as it would have been a few years ago. Time sharing and remote access on large computers has passed the experimental stage and is likely to be widely available; this means that a single computer, holding the citation index in its memory, can serve a number of stations a few hundred miles away, at low cost. All the equipment the user needs is a teletypewriter or similar terminal device.

The arrangement of the index is of great importance. In the case of a printed index, searches which do not correspond to the arrangement of the index are entirely impractical; in SCI in particular, the name of the (first) author of the cited paper must be known. (The publisher of SCI has a machine-readable record of the index and could presumably perform other searches, but only at high cost.) In TIP, a variety of searches is possible, but not all are equally easy. The file is well arranged for finding (citing) papers in one or more given journals or volumes. To find all citations to a given paper, no matter when or where (within the coverage of TIP) they appear, is more costly but by no means prohibitive. Such a search would be made easier by an inverted file; indeed, such a file could be produced easily from the main TIP file, but at present the cost of the added storage would exceed the cost of computer time saved. It is just one of the advantages of machine-readable files that they admit different types of search questions as well as reordering of the entire file, and printing of various selections from the file in various arrangements.

The user of a printed index enjoys the benefit of seeing additional related information contiguous to the entry being searched. He incurs the trouble of having to copy his answers. The computer provides printed copy, but in the case of time-shared computers the user loses time while waiting for the teletype machine to print answers. Additional waiting time can occur when the computer gets too many simultaneous inquiries.

5. Plan of Experiments

The arguments outlined above led to the conviction that citation indexing should be done by computer if at all. A series of experiments was first planned in 1964 for the purpose of becoming familiar with this technique and judging its merits. At that time the TIP file at MIT was the only conveniently accessible citation index in computer-readable form. Being limited to physics, it seemed a logical choice for experiments to be conducted by the National Bureau of Standards.

The set of *citing* papers in the TIP file consists of all papers in a small selected group of physics journals, numbering about 20 in 1964 and since expanded to over 30. It includes all of the most important physics journals in the United States, and a sampling of foreign journals. In the case of Russian journals the English translation is used. For most of these journals, all volumes beginning with 1963 are included. Table 1 presents a recent listing of the TIP "library" of citing journals.

The set of *cited* papers consists of most, but not all, the papers cited in the citing papers. Included are those, and only those, papers appearing in a list of about 250 journals. There is no time limit; all volumes are covered. All citing journals, of course, are among the cited journals, including the volumes earlier than 1963. Excluded from the list of cited papers are all references to journals other than the 250 or so "cited journals" as well as all references to the nonperiodical literature.

In addition to the file itself, TIP offers a computer program for retrieving information from the file. The file and the program are part of the CTSS time-sharing system at Project MAC [5]. The program can retrieve items according to several criteria, such as papers by a given author, papers containing certain expressions in their title; the two criteria most important for our purposes are

(a) finding citations to a given paper,

(b) finding papers which have at least one citation in common with a given paper.

We had initially planned to concentrate on experiments with (b); gradually we had to recognize the need for conducting a large part of the investigation by means of (a). Obviously (b), which is known as "bibliographic coupling" or "share bibliography" searching, is a special case of (a); a request of type (b) can be executed by finding all citations in the given paper and then finding citations *to* any of them. In (a) the given paper must be in one of the 250 "cited journals," in (b) it must be in the "library," i.e., it must be in one of the 30 "citing journals" since 1963.

In the spring of 1964 we began by enlisting the participation of a number of physicists at the National Bureau of Standards. There were seven individuals, plus one close-knit group of several people. They represent a cross section of research in physics at the Bureau, are active in creative work and at the same time high enough in the administration so that their judgment of our experiment would carry weight. We attempted to make the least possible demands on the participants' time. Each of these persons supplied a list of his publications. Their size may be seen in table 1. Our first intent was to attempt bibliographic coupling, i.e., search for papers which share references with those given, separately for each of the eight bibliographies.

Since at that time long-distance access to the MAC computer was not practical, the original plan was to have the computer runs carried out by the TIP group at MIT. After some time it became clear that this would not be feasible, and we began to look for remote access to the computer.

Table 1

LIBRARY

THE DATE IS 01/17/67
THE VOLUMES AVIALABLE TODAY ARE . . .

ANNALS OF PHYSICS
J384–ANNPHY–ANN PHYS
V 26–39
APPLIED PHYSICS LETTERS
J646–APPLET–APPL PHYS LETTERS
V 3–9
CANADIAN JOURNAL OF PHYSICS
J55–PHYCAN–CAN J PHYS
V 42–44
DISCUSSIONS OF THE FARADAY SOCIETY
J153–DISFAR
V 40
HELVETICA PHYSICA ACTA
J43–PHYHEL–HELV PHYS ACTA
V 37–39
INDIAN JOURNAL OF PHYSICS
J164–INDJPH–IND J PHYS
V 38–39
ETP LETTERS
J821–JETLET–JETP LETTERS
V 1–4
JAPANESE JOURNAL OF APPLIED PHYSICS
J612–PHAPJA–JAPAN J APPL PHYS
V 3–5
JOURNAL OF APPLIED PHYSICS
J11–PHYAPP–J APPL PHYS
V 35–37
JOURNAL OF CHEMICAL PHYSICS
J12–JCHEPH–J CHEM PHYS
V 40–45
JOURNAL OF CHEMICAL PHYSICS SUPPLEMENT
J842–JCHEPS
V 43
JOURNAL OF MATHEMATICAL PHYSICS
J227–MATHPH–J MATH PHYS
V 6–7
JOURNAL OF THE OPTICAL SOCIETY OF AMERICA
J45–JOPSOC
V 55–56
JOURNAL OF THE PHYSICAL SOCIETY OF JAPAN
J80–PHYSOJ–J PHYS SOC JAPAN
V 19–21
IL NUOVO CIMENTO
J17–NUOCIM–NUOVO CIMENTO
V 31–45
IL NUOVO CIMENTO SERIES B
J841–NUOCIB
V 40–42
IL NUOVO CIMENTO SUPPLEMENT
J843–NUOSUP
V 3
MOLECULAR PHYSICS
J160–MOLPHY
V 9–11
NUCLEAR PHYSICS
J682–NUCPHY–NUC PHYS
V 50–85
PHILOSOPHICAL MAGAZINE
J28–FILMAG
V 9–14
PHYSICA
J21–HYSICA–PHYSICA
V 30–32
PHYSICS LETTERS
J49–PHYLET–PHYS LETTERS
V 8–23
THE PHYSICS OF FLUIDS
J799–PHYFLU–PHYS FLUIDS
V 7–9
PHYSICAL REVIEW
J1–PHYREV–PHYS REV
V 133–150
PHYSICAL REVIEW LETTERS
J41–PHYRET–PHYS REV LETTERS
V 12–17
PHYSICAL REVIEW, SERIES B
J199–PHYREB–PHYS REV B
V 133–140
PROCEEDINGS OF THE PHYSICAL SOCIETY (LONDON)
J3–PHYPRO–PROC PHYS SOC
V 83–89
PROCEEDINGS OF THE ROYAL SOCIETY
J23–PROCSO
V 283–294
PROGRESS OF THEORETICAL PHYSICS (KYOTO)
J29–PROPHJ–PROGR THEORET PHYS
V 31–36
PROGRESS OF THEORETICAL PHYSICS SUPPLEMENT
J840–PROPHS
V 34–965
SOVIET JOURNAL OF NUCLEAR PHYSICS
J825–SOVJNP–SOVIET J NUC PHYS
V 1–3
SOVIET PHYSICS–JETP
J669–SPJETP–SOVIET PHYS JETP
V 18–23
SOVIET PHYSICS–SOLID STATE
J310–SPSOLS–SOVIET PHYS SOLID STATE
V 6–8
SOVIET PHYSICS–TECHNICAL PHYSICS
J790–SPTPHY–SOVIET PHYS TECH PHYS
V 9–11

END OF LIBRARY

Shortly thereafter the TIP file was temporarily removed from the MAC computer to allow for remodeling the system. It was not until November of 1965 that we could run the first trials from remote consoles at NBS in Gaithersburg, Md. After an interruption of a few months these experiments were resumed in the summer of 1966, and a number of trial runs have been made since then.

Meanwhile our bibliographies were 1½ years old and contained few items of 1963 or later; citing papers published before 1963, originally included in the TIP file, had been removed from the file. A meaningful search for shared references was possible for only one author, whose bibliography included 7 papers in 1963 and 1964 in one library journal. We searched the last three years of the same journal for shared references with one of these and found 20 papers. The author decided (by inspection of the titles and authors' names) that about 13 of the 20 would be relevant to his field of interest—a high percentage, considering that any paper was "found" if it had even one reference in common with one of his own papers. It would have been interesting to see whether and how much the "relevance ratio" improves if the search is limited to papers sharing 2, 3, . . . references with the source item; but this would require a larger sample, and programming of special instructions in CTSS outside the TIP system. Nor could we obtain any good information on the "recall ratio" i.e., on how many of the potentially relevant papers were found and how many were missed by the search. Before we had a chance to extend the search to all seven eligible papers and to additional journals, our access to the computer was temporarily interrupted.

When we were able to resume work in the summer of 1966, we decided to follow two lines of approach. One was to search systematically for citations *to* the eight sample bibliographies, in order to obtain comparable data for different areas of physics and some feel for the relevance of references so obtained. The other was to use any handy opportunity to explore the degree of completeness of literature lists obtained through citation indexing, perhaps iterated, and bibliographic coupling; such opportunities are infrequent and have to be used as they arise. A hoped-for by-product of both approaches was to gain experience with remote computer access.

6. Relevance

Some facts about the eight bibliographies of NBS authors used in this study are summarized in table 1. There were a total of 363 references, of which 297 were to publications in 1940 or later. Of this latter group, 176 were to papers in the 250 "cited journals" covered by TIP. The other references were either to papers in other journals or to non-journal items (books, reports, proceedings, etc.). The TIP library, i.e., the set of citing papers recorded in TIP, was searched for references to each of the 8 groups of papers making up this list of 176. At the time of these searches the TIP library covered about 25 journals mostly for the years 1963–1965, in some cases also part of 1966. As might be expected, the results vary greatly from author to author. The TIP library consists mostly of journals heavily oriented toward modern fields of physics, such as nuclear and solid state physics. This is one reason for the relatively high number of citations to the papers of authors B, C, E, and H.

Other factors come into the picture too. The absence of citations to papers by author F reflects not only the fact that his fields of interest are not well covered by the TIP library, but also the fact that they are older than the others; 7 of the 11 papers date from before 1950. It is known from other studies of citation indexing that most citations come in the first five years after the appearance of the cited paper.

The purpose of this group of searches was to judge the relevance of literature obtained by citation indexing. No precise results could be expected, because the concept of relevance is so vague. Thus there seemed to be no need to spend great effort in determining the degree of relevance of retrieved papers, especially in view of our desire to minimize the inconvenience to participating authors. These authors were merely shown the retrieved lists (titles and authors of citing papers) and were asked for a quick judgment of whether these papers were pertinent to their field of interest. The answer was affirmative in a large majority of cases, as indicated in the last column of table 2. Uncertainties arose rarely from inability to judge the content of a retrieved paper on the basis of its title and author(s) alone; they arose commonly from judgments like "moderately interesting" or "slightly interesting" and from the lack of consistency among the judging authors in using such terms.

The retrieved references include some papers by the cited authors themselves, in cases where they publish in one of the 25 journals of the TIP library and reference their own earlier papers. This is not a major factor; there is only one such paper by author B, 4 by author C, 3 by D and 6 by the group of authors represented by H (including, of course, mutual references of these authors to each other).

Since all authors are affiliated with the National Bureau of Standards, it may be remarked parenthetically that the NBS Journal of Research is not among the 250 "cited journals" of TIP; thus all papers published by the authors in that journal had to be omitted from the study. Another

TABLE 2. *Citation searches to publications of eight solicited authors*

Author	Field	Number of items in bibliography since 1940	Number in cited file	Number retrieved in citation search	Relevant
A.	Cryogenics	52	31	11	7
B.	Photonuclear reactions	29	18	44	Many
C.	Molecular spectroscopy and statistical mechanics.	53	43	53	50
D.	Molecular spectroscopy	49	47	68	60
E.	Neutron reactions	20	[c] 6	9	9
F.	Metrology and geophysics	[b] 58	11	0	.....
G.	Metrology and electricity	8	2	1	0
H [a]	Solid state physics	28	18	43	Many
	Total	297	176	229	.....

[a] Group bibliography.

[b] Does not include 80 publications prior to 1940.

[c] Excludes a number of older papers in "cited file" but omitted from search because somewhat different field.

curiosity concerns one paper retrieved among the references to author A, which the author judged totally irrelevant; a follow-up showed that the paper made reference, not to a paper by author A, but to another paper starting on the same page in the Physical Review as one of author A's papers. Such papers are indistinguishable in the TIP system if they occur only as cited papers; this is only a minor annoyance here, since it causes only a small number of false drops; it would be more serious in an index including letters to the editor and other short items.

7. Completeness

It is quite difficult to determine how completely a procedure—for instance a particular request to a citation index—recalls all of the desired literature.[3] The difficulty lies in finding a control procedure which can be relied upon to produce complete recall. Usually a painstaking literature search by a subject matter specialist is the only method. We availed ourselves of a few opportunities in which such searches had been made or in which we could persuade a scientist to make them. In some cases it turned out after the fact that, for some reason or other, no valid inferences could be drawn; we shall nevertheless mention in the following even these unsuccessful attempts.

(a) In our first trial we were hoping, while examining recall, also to explore the usefulness of citation indexing for data retrieval—the primary problem which has caused us at NBS to take an interest in this subject. We chose NBS Circular 500, "Selected Values of Chemical Thermodynamic Properties," first published in 1952 and now coming out in a new edition; more specifically we selected the sections dealing with the elements Si, Ge, Sn, Pb, Ga, In, Tl (and their compounds with those elements preceding them in the standard order of arrangement). The first edition contains an extensive bibliography for each of these seven groups. It was our plan to look for citations to items in this older bibliography, and compare the list so obtained with the list of recent publications actually used by the compilers of Circular 500 in the new edition.

In the first edition of Circular 500, there were about 270 references in the seven sections included in our study, but only 123 of them were to the 250 "cited journals" of TIP; the others were references either to the nonperiodical literature, or to journals not included in the TIP list of cited journals. We then searched the entire TIP file of 25 citing journals for papers citing any of

[3] The term "recall ratio" is commonly used for the ratio between the number of items retrieved by the tested procedure to the total number of items that one would wish to retrieve. This is in contradistinction to the "relevance ratio" discussed in the preceding section.

these 123 papers. (This required some maneuvering with tapes and specially created files; the TIP System does not easily accommodate searches through many journals for finding references to many different papers, for customers having only a small amount of computer storage allotted for their use.) In the end, somewhat to our surprise, we found only six such references. In retrospect the reason for this disappointing result seems to be that the subject of Circular 500—thermodynamic properties of materials—is not usually represented in the citing journals covered by TIP. The latter focus on the mainstream of modern physics. Recent papers on the more classical subjects, such as thermodynamics, would have to be looked for in a different set of journals.

(b) We then turned our attention to a field which is strongly represented in TIP, nuclear physics. From a nuclear physicist at NBS ("author J") we obtained five "fundamental references" on the subject of neutron scattering. They are:

Phys. Rev. **55,** 190, W. E. Lamb, Jr., Capture of neutrons by atoms in a crystal (1939).

Phys. Rev. **94,** 1228, G. C. Wick, The scattering of neutrons by systems containing light nuclei (1954).

Phys. Rev. **95,** 249, L. Hove, Correlations in space and time and Born approximation scattering systems of interacting particles (1954).

Phys. Rev. **101,** 118, A. C. Zemach and R. J. Glauber, Dynamics of neutron scattering by molecules (1956).

Phys. Rev. **110,** 999, G. H. Vineyard, Scattering of slow neutrons by a liquid (1958).

Searching the entire TIP library for papers citing one of the five basic ones, we obtained 93 papers. By inspecting the titles and authors' names of the referenced papers, author J concluded that all but about 15 were relevant. To keep further work within manageable limits, he selected 16 out of the 78 relevant papers; these pertain to a narrower subject, broadening of neutron resonance lines. With these we searched for papers with shared references, limiting the search to five Soviet journals (a total of 15 volumes). This resulted in a list of 80 papers.

As was to be expected, a majority of them were not relevant; our aim was to get high recall, even at the expense of high admixture of irrelevant papers.

To test the completeness of recall, author J kindly examined two of the 15 volumes searched, namely, volumes 18 and 20 of the English translation of Soviet Physics—Journal of Experimental and Theoretical Physics (JETP). Of the 80 papers found in the TIP search, 22 were in these volumes, including nine relevant, one marginal, 12 not relevant to the topic (broadening of resonance lines). By looking only at the Tables of Contents of the two volumes (i.e., authors and titles), author J spotted 19 potentially relevant papers. Closer examination of the papers themselves showed that six of the 19 were not relevant and three were marginal, leaving only 10 clearly relevant. That is to say, the precision of a human search using authors and titles only was in this instance not much better that of the TIP search. Even more surprising was the low degree of overlap between the two searches. Only six papers appeared in both searches; five of these were relevant ones; one was irrelevant, despite the fact that its title had looked pertinent and that it had one citation in common with each of two previously found papers in the field. The following table summarizes the result.

	Search for papers on neutron resonance broadening				
	Found by author J			Not found by author J	Total
	Relevant	Marginal	Not rel.		
Found by TIP: Relevant	5	*	*	4	9
Found by TIP: Marginal	*	0	*	1	1
Found by TIP: Not rel.	*	*	1	11	12
Not found by TIP	5	3	5	?	13+
Total	10	3	6	16+	

Not only was the relevance ratio of author J not nuch better than that of the TIP search, but his recall ratio was not much better; he missed four relevant and one marginal paper found by TIP, while TIP missed five relevant and three marginal papers found by author J. It is true that the omissions of author J were not entirely due to misleading titles, but were in part caused by the rapidity of scanning the Tables of Contents; but it is legitimate to charge the manual search method with being tiring and therefore conducive to error.

(c) In section 6 we mentioned 44 papers retrieved in a search for citations to the publications of author B. It happens that this author maintains a bibliography intended to be reasonably complete on photonuclear reactions and related phenomena. We recognized that this bibliography could serve as a control for our purposes without requiring any great additional effort on the part of the author. To investigate how quickly we could advance toward a complete bibliography, we decided to continue from the 44 citing papers retrieved earlier, by searching the library for papers which share at least one reference with one of them. This is equivalent to obtaining all papers cited by the 44, and then obtain further citations to these cited papers. (Because of minor technicalities three of the 44 papers were omitted from the search.)[4]

The author's file consists almost entirely of papers from five journals—*Nuclear Physics, Physical Review, Phys. Rev. Letters, Physics Letters*, and *Soviet Physics—JETP*. At the time of this study there were 235 such papers. In addition there were 25 papers scattered among other journals, and 12 non-journal items. Of the 235, only 69 were from 1963 or later, the others were thus too old to be retrieved from the TIP library. Our plan was to see how completely these 69 papers could be retrieved by a few steps of citation indexing. Table 3 shows the results of the search. To save time, only those volumes of the five journals were searched in the TIP library which contained articles on author B's list. These volumes contained a total of 5,845 papers, of which 1,074 were "retrieved" by the criterion that they share at least one reference with one of 41 papers citing author B. Most of these 1,074 were not relevant to photonuclear reactions; but included among them were 57 which were relevant, as evidenced by the fact that they were listed in author B's bibliographic file. Thus 57 out of 69 possible papers were retrieved in one step of bibliographic coupling, starting from a base of 41 papers which undoubtedly were not an optimal base.

Having thus learned that bibliographic coupling with these 41 papers was sufficient to retrieve most papers of interest, we applied the same search to a number of other journals in the TIP library; journals which author B was not likely to consult because they do not emphasize nuclear physics. The results are shown in table 3b. Of the less than 5,000 papers in these journals, 259 satisfied the search criterion. Again most of them were not relevant, but author B found three papers among them which were of interest, and which had not previously come to his attention. This points up one of the more promising applications of citation indexing: to search "low-yield journals" which are not likely to contain many items if interest, which an author would therefore not normally look at, and select from them a manageable list of papers which still contains most of the interesting items.

(d) Author E, in addition to supplying the references listed in table 2, had also given us, at our request, a list of five papers by other authors, which he considered as characteristic of his field of interest. These were:

Fluctuations in partial radiation widths of U^{239}, H. E. Jackson, Phys. Rev. **134,** B931.

Neutron resonance spectroscopy III Th^{232} U^{239}, J. B. Garg, J. Rainwater, J. S. Petersen, and W. H. Havens, Jr., Phys. Rev. **134,** B985.

Theory of radiative capture in the resonance region, A. M. Lane and J. E. Lynn, Nuclear Physics **17,** 563 (1960).

Anomalous radiative capture, Lane and Lynn, Nuclear Physics **17,** 586 (1960).

Parameters and gamma ray spectra, R. T. Carpenter and L. M. Bollinger, Nuclear Physics **21,** 66 (1960).

[4] This search would have been rather slow if done by remote access. The authors thank Mr. W. Mathews of MIT for running it on the computer.

TABLE 3. *Search for papers sharing citations with any of 41 papers citing author B*

a. Journals previously scanned by author

Journal	Volumes searched	Number of papers			
		Searched	Retrieved	Relevant	Known to author
Phys. Rev	B133–B140	1,785	474	7	8
Phys. Rev	141, 143, 149	577	73	7	7
Nuclear Phys	51–54, 56–57, 59–60, 63–64, 70, 74–76.	781	292	29	35
Sov. Phys. JETP	19	266	11	0	2
Phys. Ltrs	8–9, 11–13, 16–19	1,788	183	12	14
Phys. Rev. Ltrs	12, 15	648	41	2	3
	Total	5,845	1,074	57	69

b. Low-yield journals

Journal	Volumes searched	Searched	Retrieved	Relevant	Known to author
Appl. Phys. Ltrs	3–9	899	0		
Melv. Phys. Acta	37–39	238	14		
Prog. Th. Phys. (Kyoto)	31–36	400	40		
Nuovo Cim	31–45	1,151	65		
Canad. J. Phys	42–44	476	30		
Physica	30–32	437	15		
Sov. J. Nuc. Phys	1–3	462	64		
Proc. Phys. Soc. Lond	83–89	765	31		
	Total	4,828	259	3	0

A search of the TIP library for citations to these papers brought forth eight papers, including three in Russian journals. All were judged highly relevant to the field, and the discovery of the Russian papers was a welcome bonus to the author, but the small volume of the result was disappointing.

(e) A physicist at the National Bureau of Standards ("author K") needed a bibliography on computation of molecular wave functions, and was persuaded to attempt use of the TIP file in parallel with a conventional approach. He supplied a list of 33 authors and of the journals in which they habitually publish—two or three for each author, mostly, *J. Chem. Phys.* and *J. Mol. Phys.* A search of these journals in TIP, specifying the authors, gave 133 papers. The same search was performed by an assistant reading through the tables of contents of the journals. Comparison of the results showed that (1) the search in TIP missed three papers, two because the author's name was misspelled in the computer file and one because of a confusion on input; (2) the human assistant overlooked several papers; (3) the 133 papers produced by TIP included about 20 by authors with names similar to, but not identical with, those requested (usually different initials), because the input was not specific enough. This, of course, is only a minor annoyance, less serious than omission of a desired reference.

We then searched the TIP record of *J. Chem. Phys.* and *Mol. Phys.* for citations to any of 129 papers found in the first step, specifically, all those of the 133 which appeared in these two journals, omitting only four papers from other journals. The 20 or so false drops of the first step were included because they had not been discovered at the time of this search. The result was a list of 167 papers, of which 99 were judged relevant by author K.

(f) The search conducted for author K gave as a byproduct a comparison between manual and computer searching. Another experiment was made expressly for this purpose; it consisted in duplicating by means of the Science Citation Index [4] one of the searches previously made in TIP. We caution against considering this as a comparison between SCI and TIP for completeness; a meaningful comparison for this purpose would require a far larger sample than the one used here. We were interested in ease and convenience of the processes, and these can be judged to some extent even from a single attempt. As it turned out, however, the experiment revealed more about peculiarities of the two indexes used than about the difference between manual and machine methods in general.

The search for references to author B, reported in section 6 above, started with 29 publications by B and resulted in 44 citations to them. Of these, 26 appeared in 1964. We searched manually in SCI for 1964 and found 25 references to B. Of the two sets of references, 21 overlapped. SCI missed five because the citing journals (*Phys. Letters, Sov. J. Exp. & Theor. Phys.*) are not covered by SCI. TIP missed four citations, of which one is to a cited journal not covered by TIP (*J. Res. Nat. Bur. Standards*), three are to citing journals not in the TIP Library (*Rev. Mod. Phys., Zeitschr. f. Phys., Atom. En. Rev.*). As for convenience, compare (1) walking to the library in adjacent building, starting to search, finding an unanticipated need for additional information, having to walk back to one's office to locate it, and then once more to the library; and (2) walking down two flights of stairs to the nearest teletype, dialing the computer, finding all lines busy, repeating this a little later, and finally deciding to run the search late at night when the computer is less in demand. Once connection is made, the computer needs only a few minutes of main frame time, but these are spread over 1–2 hours of real time; at the end of this time we have a neat printed list of all desired citations, 1963–1965. With SCI, we quickly located author B in the 1964 volumes and copied 18 citations; realized that this was incomplete; remembered that SCI lists only first authors; fetched B's list of publications; located all papers where B appears as second author; noted the first author (mostly the same on all papers—call him L); located author L in the 1964 SCI; picked out, with some effort, citations to those of L's papers on which B was co-author (this requires going back to B's publications list each time, since B's name does not appear in SCI with these papers) and finally copying seven references to these papers. After spending about $1\frac{1}{2}$ hours in the library, we now had the 1964 citations and not enough patience to repeat the process for other years. Comparison of the two sets of citations revealed that they were completely explained by the different coverage of the two indexes, so that there appeared to be no erroneous omissions.

8. Conclusions for the User

What does the user get out of citation searches? It depends on what he has got out of other search methods, on his field of interest, and on the purpose of his search.

A first fact which emerges from our experiments is that papers which cite a given author are usually in the author's own field of work, more so than one might have expected. Papers in other fields are a small minority. Many of the papers found in these searches were already known to the cited authors, but most authors found at least a few papers—typically perhaps 25 percent—which were new to them. Authors who keep reasonably complete bibliographies in their fields are not likely to find much of value in such a search; in our sample, author B was in this class. But an author who, like most scientists, habitually scans only a few journals in his field and otherwise relies on personal contacts for keeping informed, is likely to find a number of references that are new to him.

The results of such a search for citations to one author's papers are quite incomplete; that is, there are usually many papers in the literature which are of interest to the author even though they do not cite his writings. One can obtain more complete coverage by an iterative process, in particular, by searching for papers which share references with papers in a given list. (This list may consist of the author's papers, or better, may include some or all of the references cited in his own paper.) Here the results will depend strongly on the size of the starting list; in our sample, starting with half a dozen papers gives few results, starting with about 40 gives over 80 percent coverage. This high coverage is bought at the price of a very large admixture of irrelevant papers. The author has to plow through a large number of retrieved papers to find the few that are of interest to him. Nevertheless the process is not entirely useless. Even if the retrieved papers constitute, say, 15 percent of the search range (as they do in table 3), the author's work in looking for relevant papers has been reduced by 85 percent. The method appears especially promising if applied to "low-yield" journals, i.e., those in which only a small fraction of all papers is in the author's field, and which the author would therefore normally not consult.

Both kinds of searches seem to be successful enough to be used as supplements to other methods of literature retrieval, but not good enough to be relied upon as the sole means, or even a principal means, by which scientists are kept informed.

It is plausible that the performance of shared-reference searches could be improved by enlarging the starting list of papers and at the same time tightening the search conditions to require two or more shared references. Unfortunately it was not possible to explore this conjecture in the course of our program. In the TIP system such a search must be conducted by first searching for all shared references and then weeding out all papers with only one shared reference, so as to retain only those with multiple shared references. This would have to be done with a large list of starting papers, but even the 41 papers of section 7c above were almost beyond the limits of machine time and memory available to us.

Sitting at a console, keying questions into it and seeing results typed out has a number of advantages which are worth listing here. We were rarely able to anticipate the volume of output generated by a question. It is well to feel one's way, search one volume at first, gradually increase the search range. When finding high yields in some journals, low ones in others, one can then combine several low-yield journals into one search, subdivide a single high-yield journal into several searches. Similarly, it is economical to combine a number of questions into a single search, to save computer time. It is then better to store the answers and print them on a subsequent run, rather than print at once, in order to separate the answers pertaining to the different questions. But this runs into the limitation of memory available to individual users of the MAC computer, and therefore is again best done by on-line operation, where one can stop just before having exhausted one's memory quota. After much experience, these precautions could be anticipated and incorporated into batch-processed computer runs, but many scientists are likely to use citation searching only occasionally, without ever becoming expert at it.

In our experience, in one hour of time spent at the console one gets anywhere from less than 1 to 10 or 15 minutes of main frame computing time; the average might be 2 minutes.[5] The commercial value of 2 minutes on a large computer is about $20, the cost of a one-hour telephone or teletype connection is of the same order of magnitude. We believe that the greater convenience and efficiency afforded by on-line operation is well worth the added cost.

We have already mentioned certain limitations which are peculiar to the TIP file, rather than to the method in general. The file covers some areas of physics poorly, other areas very well, but not completely. In time it covers three or four years. It omits all citations to the non-journal literature and to some less frequently used journals. The first of these limitations was deliberately imposed because of the experimental nature of the entire TIP project. The second is due to memory limitation of the MAC computer. The third is caused merely by the encoding method chosen—three decimal digits to represent a journal—and could be most easily overcome.

Despite the limited file size, the time taken for searching is fairly long. The TIP manual states that one search of the entire file takes three minutes of main frame time. Since that manual was written, the file has been enlarged, so that the search time has become longer.

In a dozen searches which we ran in different parts of the file, the computer searched an average of 100 papers per second; the speeds ranged from 40 to 160 papers per second, with a

[5] These numbers should not be generalized without scrutiny; they are peculiar to the time-sharing system used, and especially to its scheduling algorithm.

standard deviation (rms) of 40. It would probably take 5 to 6 minutes of main frame time to search through the present file for a single question, at a cost (if done commercially) of $50 to 75, plus a similar amount for phone time if done at long distance. Furthermore, we have found that almost invariably the machine time is somewhat longer than the nominal time given by the TIP manual, presumably because of some housekeeping operations. Finally, the nominal time is for single questions. It increases only slightly if a few questions are processed simultaneously, but it becomes several times as long for batches of 40 or 50 questions. It seems likely to us that the computer time could be substantially reduced if search programs were written with this goal in mind; as is appropriate for a research project, the original search programs were written for flexibility rather than for economy of computer time. Further time savings could undoubtedly be realized if the file were kept in several different arrangements, including one sorted on cited papers like the ISI Science Citation Index. But whether the resulting economies would offset the increased cost of storage is a question we are unable to answer.

9. Some Extrapolations

In describing our experiments with the TIP system we have noted the usefulness of the CTSS remote access time sharing system in which TIP is imbedded. The usefulness of the TIP system as one for manipulating citation index data is considerably enhanced by the properties of the CTSS system. However, in a broader sense we may view the set of experiments above as a special case of a more elaborate system for manipulating scientific information, whose eventual development we may anticipate by extrapolating from our experience with the experiments described above.

It should be noted at the outset that the notion of a citation index is itself a derivative from the historical development of conventional publication media. If we are willing, however, to accept the technological fact that scientific information communication can be mediated by a remote access computer then it becomes possible to consider a whole new class of services that such a remote access computer mediated communications system can achieve. The main arguments against such services are economic in nature. As to whether or not such eventualities will actually occur—whether in the near or far future—is more a question of economics than of technological possibility.

One such possibility which we have already explored and mentioned above is the rather straightforward one of communicating conventional mail type of information. In a system such as CTSS, it is possible to compose memoranda and to mail them to other users. The other users receive a notice of the existence of such mail whenever they enter the system, usually for other computing purposes. A modest extension of this capability allows the initiative to be taken by the system to call the user and notify him of the existence of mail. A still further extension of such capability allows its use as a distribution system for documents like reports and others of a nature less formal than conventional publications. Such a distribution system is of course entirely under the control of the mediating computer programs. Thus the possibility of differential distribution lists as a function of time and subject content become possible.

A much more important possibility is that of differential distribution of documents through a remote computer system in which the change from one time to the next is in the contents of the documents themselves rather than in the distribution list for the documents. A document, for example, may change in time as a function of use, because of the addition of editorial corrections, or the addition of comments that may be considered the equivalent of marginal notes by readers. In certain special cases these may be allowed to update the master document. Again access to the updating function to the document itself may be controlled by the author or some more complicated agency, under his control. The most important point to note here is that the contents of the document being manipulated within such a system are subject to extensive change by a network of readers and potential readers whose changes are themselves partly under the control of the author of the document.

On the question of economics of such a system the only observation worth mentioning at this point is that such a system cannot achieve practicality until the number of interconnected users becomes sufficiently large that the set of potential readers of documents may be considered to be largely included within the set of users of the system. For specialized classes of users this technological possibility is perhaps economically realizable now. For general scientific documents this of course is not yet the case.

We have thus far unnecessarily been assuming implicitly that the information to be communicated within such a remote access distributed network of users is like conventional documents. But even if we wish to preserve the archival function of conventional publication media we need neither preserve the passive nature of the documents nor their relative immutability. Our notion of a less passive more extended document might include one containing the description of procedures, like computer programs or process control procedures for example, which need not be read in a conventional sense at all but merely called by other programs and used, without the occurrence of an intermediate process of reading and understanding, by some individual person. Programs in a system like the CTSS system are conventionally used in just such a fashion. Most of the programs that a user invokes he never reads. Instead he obtains some form of certification (in this case usually informal) as to the validity of the process which he is using, and then proceeds without further inspection of the document at all.

Thus "quasi-documents" in such a system are active in the sense that they directly influence subsequent processing of information without the intermediating process of human consumption of the quasi-documents. Such quasi-documents are also less immutable as a function of time. The situation can be so arranged that when one calls upon a quasi-document for inspection only the latest version will be furnished by the system. The current version distributed will be the latest one with all the changes up till the point of calling of the quasi-document from the system. The historian or others interested in the archives may in such a case generally get access to original documents or prior versions of such a quasi-document where the archiving function exploits the variable kinds of access in such a system, older documents having been superseded by archiving them in less accessible kinds of storage like magnetic tape.

Since we are willing in these speculative comments to consider the possibility of extensive modification of the usual scientific publication process we might also briefly consider the possibility of largely avoiding this process altogether in certain cases. We have in mind the possibility of allowing the scientific information that ordinarily results in publication to reside implicitly in the authors of such documents without necessarily receiving any external form until called upon for use by the system. The way this could be achieved is in a very large system with users who may be called upon, on the initiative of the system, to serve its purposes of information access. Users whose interests and competence profiles are stored in such a system might be addressed by the system when it needs information from those users of appropriate expertise who are in consultant mode at that moment. The "consultants" would "publish" their information when it is needed and otherwise it would not be recorded. One can imagine such a system responding rapidly to scientific information needs of the moment, with of course the concomitant disadvantage of the loss to the publication process of those kinds of scientific information which anticipate future needs rather than respond to them. The solution to this problem is of course to include both possibilities within such a system. But for the evanescent needs for certain kinds of scientific information the possibility of a remote access system being helped by consultants on line who supply information when it is needed, by the machine, should not be entirely ignored.

Notice that such a system is the dual of a conventional time sharing system, which we might more profitably view as a machine information distribution system, aided by people, rather than people aided by such a machine. Such a "man-aided computer" can offer some of the advantages in use of human efforts that a conventional time sharing system offers in the utilization of machine capabilities. We are thus considering the more complex kinds of scheduling algorithms for people and their scientifically productive efforts, that might be possible in such a system, in which access to people and their information is mediated by a computer with its elaborate scheduling capabilities, just as now we have access to the computer mediated by that same kind of scheduling capability.

The above speculative comments have been inspired by the existence within the MAC CTSS system of nascent versions of most of these proposals. We have raised the issues largely because in most of these areas no technological limitation prevents us from exploiting these possibilities. Rather the demand by a large set of users for such capabilities at a level of economic feasibility is what is more likely to influence the ultimate existence of such capabilities, within a computer mediated scientific information system. Thus if potential users can begin thinking about such possibilities the ultimate economic realization of them can be speeded up.

10. References

[1] Brown, Sanborn C., A bibliographic search by computer. Physics Today, **19,** No. 5, 59 (May 1966).

[2] Kessler, M. M., The M.I.T. Technical Information Project, Physics Today, **18,** No. 3, 28 (March 1965).

[3] Pyke, Thomas N., Jr., Time-shared computer systems. Advances in Computers, **8,** 1 (Academic Press, New York).

[4] Garfield, E., Science Citation Index—a new dimension in indexing, Science, **144,** No. 3619, 649 (May 8, 1964).

[5] Corbató, F. J., et al., The Compatible Time-Sharing System (MIT Press, Cambridge, Mass., 1963).

[6] Garfield, E., and I. H. Scher, ASCA (Automatic Subject Citation Alert), A new personalized current awareness service. Amer. Behavioral Scientist, **10,** issue 5, p. 29 (1967).

(Paper 72B1–260)