Information Retrieval on a High-Speed Computer

A. R. BARTON†, V. L. SCHATZ†, AND L. N. CAPLAN†

This discussion concerns a mechanized information retrieval system for the technical library of General Electric's Aircraft Gas Turbine Division. The system is confined to the technical reports and papers available in the Division's Library. Textbooks have not been included as of yet since they are carried on a Library of Congress system and did not readily fit into the manual scheme which was developed. This discussion will be in two parts. The first part will cover the information retrieval system prior to mechanization, and the second part will cover the mechanization of this system.

The technical library was established in 1953 using a uniterm or key word coordinate indexing system. To understand this system, we will follow the progress of a publication through the various steps of the system. First, as a publication is entered into the library, it is assigned a six-digit access number (see Fig. 1). This report is typical of the technical reports generated within the Division. Another example of reports in the library would be NACA technical reports. The next step in this system is the abstracting of the document. The abstracting is done by professional librarians and then posted to a card file as shown in Fig. 2. This card is controlled according to the previously assigned access number. The next step in the system consists of reviewing the title, abstract, and document to select the most descriptive words which will identify the document. These words are primarily nouns. They become the uniterms. In our hypothetical case, these words are shown on the right in Fig. 2, just as they appear in the system.

These uniterms, along with the appropriate access number, are then posted to the uniterm file (see Fig. 3). There are 100 numbers per side of card when full. Both sides are used, and in the case of general terms such as these, they are heavily posted so that several cards are required. Certainly this system appears cumbersome at this point, but it has the advantage that, in any given technical area, the number of uniterms tends to level off at a specific number after the system is developed. In our case, this number is something under 9000. The combined system is shown schematically in Fig. 4. I think now you can see how information is recalled from this system. The requestor discusses his problem with the librarian. They decide on the uniterms to search. The librarian then furnishes the appropriate uniterm cards to the requestor. Once again referring to Fig. 3, the problem facing him can be seen. He must cross coordinate the cards to find numbers which apply to all uniterms. In our case, we have used three uniterms. A more typical case would be four to six uniterms but with less access numbers per term. In any case, if the requestor is persistent, he will come up with some of the matching numbers. The librarian can then go to the abstract file for the abstract cards. These are perused by the requestor who in turn selects from the abstracts those documents he desires to read in full. In the case of the three uniterms we have selected, each has over 1000 access numbers posted to it. It is easy to see the difficulty of cross coordinating these cards. This difficulty did little to promote the use of the technical library. The result was duplication of experiments, technical studies, etc., with the attendant delays in time and increases in cost.

Obviously, something better was needed. That "something better," we feel, was the program written for our IBM 704. This program is basically a mechanization of the manual system with very little effort to change the system itself. This program is in two parts. Part one is file updating and the cross coordinating of the master uniterm tape, or, referring to the manual system, the uniterm card file. Part two concerns the selection and printing of abstracts. (Part one can be run independently of part two if desired.)

Cross coordinating, at present, is done on a strict AND system. That is, an access number must appear under each uniterm used in the search. The possibility of an AND/OR system was considered and rejected. However, the AND/OR approach is being used on a modification of this program for a different type of information retrieval system. If no information is found, a modified list of uniterms can be developed by the requestor and the librarian, and another run made.

It was decided during the development of the program that one of the features of part one was to update new information into the file at the time of cross coordinating. This decision was based on the knowledge that in any information retrieval system there is a major problem of keeping the system updated. Thus, all uniterms with their associated access numbers were stored on the master uniterm tape in alphabetic order.

All search information and updating information can be read into the machine from either the card reader or tape. This information is processed to determine if it is in alphabetic order and if there is any updating information. If the input is not in alphabetic order, it will be sorted within the machine. If there is no updating, a new uniterm master tape will not be made up. Both of these decisions are made within the machine by interpreting the input information.

To utilize best the 32,000 locations of memory in the 704, it was decided to use long records of information.
A STUDY OF COMPRESSOR VIBRATION UNDER VARIED LOADING CONDITIONS

By

J. Doe

JET ENGINE DEPARTMENT
AIRCRAFT GAS TURBINE DIVISION
GENERAL ELECTRIC COMPANY
CINCINNATI 15, OHIO

A STUDY OF COMPRESSOR VIBRATION WHERE THE ROTOR HAS BEEN UNBALANCED AND THEN RUN UNDER VARIOUS CONDITIONS OF AIR FLOW, TEMPERATURE, TURBINE, UNBALANCE, VIBRATION

CLASSIFICATION

AGT-TECHNICAL INFORMATION SERIES

REPORT NO. M57AGT3471

A STUDY OF COMPRESSOR VIBRATION UNDER VARIED LOADING CONDITIONS

Fig. 1—Title page of AGT Report.

Fig. 2—Abstract card with uniterms.

Many of the uniterms, with only a few access numbers posted to them, are combined into one record with a total length of not over 7000 words. A total of 9000 locations is reserved for reading in these records. If any record exceeds 7000 words, the program will try to separate this record into two records. If the record exceeds 8995 words and cannot be broken down, the program will be halted. Because of this factor, no uniterm can have over 8995 access numbers posted to it. For a larger system, this could be easily modified so that there would be no limit to the number of access numbers posted to a uniterm. Another feature to help cut down on the length of records is whenever a new uniterm is added, this uniterm will start a new record. This will also tend to break one record into two records because this uniterm may fall alphabetically between two uniterms that are in one record (see Fig. 5).

As the machine does the cross coordinating, each search is stored in variable length buckets in memory. The total length of these buckets is also 9000 words. If the amount of information stored in these buckets exceeds 9000 words, the uniterm that caused the overflow
is stored in a temporary area and the search continues. As more uniterms are read and cross coordinated, the length of the buckets is decreased, permitting the addition of another search at the end of the last bucket. When the master tape has been completely read, the program rewinds the tapes and makes a second pass on the tape using the uniterms from the temporary area. At this point, let me stress that a second pass is only made if there is an overflow on the first pass (usually over 50 searches). In most instances, this will not occur.

If on the second pass all uniterms cannot be processed, the program will notify the operators that the search is too large and must be made smaller.

The present system contains about 35,000 abstracts with an average of ten uniterms each, or a total of over 350,000 numbers posted to the master uniterm tape.

Part one will handle 99 searches at once with no limit on the number of uniterms per search. It will also handle unlimited updating. A normal run is about thirty searches and updating of about 2000 access numbers into the general file. The time required is about two minutes for searching and four minutes for updating, or six minutes total. Tape assignments for part one are shown in Fig. 6.

Part two is the printing out of all abstracts which correspond to the access numbers discovered in the first part. One million abstracts have been allowed for in the program. Timing for part two is approximately four minutes per 10,000 access numbers into the general file. The time required is about two minutes for searching and four minutes for updating, or six minutes total. Tape assignments for part one are shown in Fig. 6.

Part two is the printing out of all abstracts which correspond to the access numbers discovered in the first part. One million abstracts have been allowed for in the program. Timing for part two is approximately four minutes per 10,000 abstracts. At the present time, 10,000 abstracts are placed on a master tape in numerical order by access number. A statistical study is now being conducted so that abstracts will be in order by frequency of use. This will significantly improve the timing for part two as we expand our system. Tape assignments for part two are shown in Fig. 7.

Originally it was thought that the program would, in most cases, go right on into part two. Therefore, referring to Fig. 6, it can be seen that the only tapes available at this time are tapes 7 and 8. The first abstract master to be read is on unit 7, the next on unit 8. While the program is searching tapes 7 and 8, the operator can mount other abstract master tapes on units 4, 5, and 6. After each master tape is searched, a continuous loop is set up selecting tapes 7, 8, 4, 5, 6 — 7, 8, 4, etc. The number of tapes to be read is determined by an input card. All access numbers found in part one are sorted into numerical order before starting part two.
At the beginning of each master tape is an indication of the range of access numbers of that tape. Each number that the program is looking for is compared against this range, and if it is not on this tape, an on-line comment will tell the operator we are finished with this tape. The program will then modify all addresses to pick up the next tape unit in sequence.

The output of this system is in two parts as shown in Fig. 8. At the top is a sample listing of those access numbers located by the system. Below is a set of single sheets of printing. Each sheet contains the abstract corresponding to one access number as shown in the column on the left. Each sheet is pre-addressed for direct mailing. If there is a security classification, it is shown. If for any reason it is considered necessary to suppress printing of security or proprietary information, this suppression is under the control of a sense switch. It is now possible for the requestor to review the abstracts, select those for which he would like to receive the original document, check the appropriate access number on the list on the left, and return this single sheet to the library.

I think that many of the advantages of this system are obvious, among which are speed, cost, designation of security classification, and the direct mailing feature. Advantages that are not obvious are:

1) Reduction in amount of human handling with the resultant errors.
2) All information is in narrative and is alphabetic. There is no coding.
3) Complete abstracts are readily available to the requestor.
4) The need for extensive manual files is eliminated.
5) No information ever need be removed from the system so no information can be lost.

This system has been in operation in General Electric since September, 1958. Some of the modifications that we have found to be advantageous through experience are:

1) In most cases it is desirable to stop the program after part one and examine the print-out before going on to part two.
2) It is advantageous to be able to supply to part two the ranges of access numbers wanted on each search.
3) An upper limit should be placed on the amount of abstracts to be printed if it is requested that the program continue on into part two without stopping after part one.
4) A method of combining several uniterms into one composite term on any individual search is extremely necessary.
5) If any uniterm reduces the number of access numbers found to zero, this term is eliminated and the search continues as if this word was not given in this search.

These are only a few of the conclusions we have reached. Future experience, we are sure, will dictate many other additions or deletions to this program. Of course, we are looking forward to the time when with new equipment we will be able to search tapes simultaneously, thereby reducing our running time by a factor of approximately 10 to 1.

We are presently planning to expand this system to mechanize the records involved in checking material into and out of the library. We also plan to develop an automatic overdue notice system.

This information retrieval program has enjoyed wide acceptance in our plant. We have received requests to modify the program to process various types of personnel registers, engine test data files, specialized blueprint files, and various other types of information systems. In any place where the key word coordinate indexing system or some variation of it can be used, this program seems to be the answer to many of our problems.
The Next Twenty Years in Information Retrieval: Some Goals and Predictions*

CALVIN N. MOOERS†

A HISTORICAL PERSPECTIVE

ALTHOUGH information retrieval has lately become quite a fad, I intend in this paper to stand back and take an unhurried look at what is going on, and try to predict where this field must go and what it must do in the future. "Information retrieval" is a name that I had the pleasure of coining only eight years ago, at another computer conference.1 The name has come a long way since then.

In thinking about a definition of information retrieval and in considering the future of this field, we must take an evolving view. At the present time, information retrieval is concerned with more than the mere finding and providing of documents; we are already concerned with the discovery and provision of information quite apart from its documentary form. To encompass future developments, as we shall see, even this broad view of information retrieval will have to be modified and extended.

When we speak of information retrieval, we are really thinking about the use of machines in information retrieval. The purpose of using machines here, as in other valid applications, is to give the machines some of the tasks connected with recorded information that are most burdensome and unsuited to performance by human beings. At all times, it is important to remember that it is the human customer who uses the information-retrieval system who must be served, and not the machine. It makes a difference who is served, and this little matter is sometimes forgotten in computer projects.

To get a historical perspective of the introduction of machine methods to information retrieval, let us look back over a bit of history. I think that it can be said that the introduction of machine methods has followed the realization of a need, backed by pressure and means to do something about the need. Thus, although quite powerful mechanical methods could have been developed by the technology of the Hellenistic Era for the Library of Alexandria, other methods of retrieval, presumably based upon human memory, and the making of lists, were apparently considered quite satisfactory. The simple, though powerful, mechanical technique that could have been used at Alexandria is the method of perforated stencils invented by Taylor in 1915, which has sometimes more recently been called "peek-a-boo."2

The British inventor Soper in 1920 patented a device which was an improvement upon Taylor's perforated stencils, and Soper described the use of his mechanism for information retrieval employing some truly advanced conceptions.3

Much more attention, however, has been given to the development of devices for scanning and selecting upon film strips. This work was apparently spurred by the perfection of motion picture films and cameras. Goldberg in 1931 patented one of the earliest film-scanning and photographic-copying devices.4 Independently, Davis and Draeger during 1935, in the early days of the American Documentation Institute, in connection with their pioneering work in microfilm documentation, investigated the feasibility of a microfilm scanner using a decimal coding.5 Apparently stimulated by reports of this work, V. Bush and his students at M.I.T. in 1938–1939 built perhaps the first prototype machine along these lines, a microfilm scanner with each frame of text delineated by a single decimal code for the subject, and a photoflash copying method. However, they were unable to interest any commercial or governmental organization in the device, and wartime distractions intervened soon thereafter, so the project was dropped. Two more "rapid selectors" based upon these same general principles have been built,6,7 but for various reasons neither of them has operated in a fashion that is consid-

* This work has been supported in part by the U. S. Air Force Office of Sci. Res. through Contract No. AF 49(638)-376. All opinions are those of the author.

† Zator Co., Cambridge, Mass.
