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.)

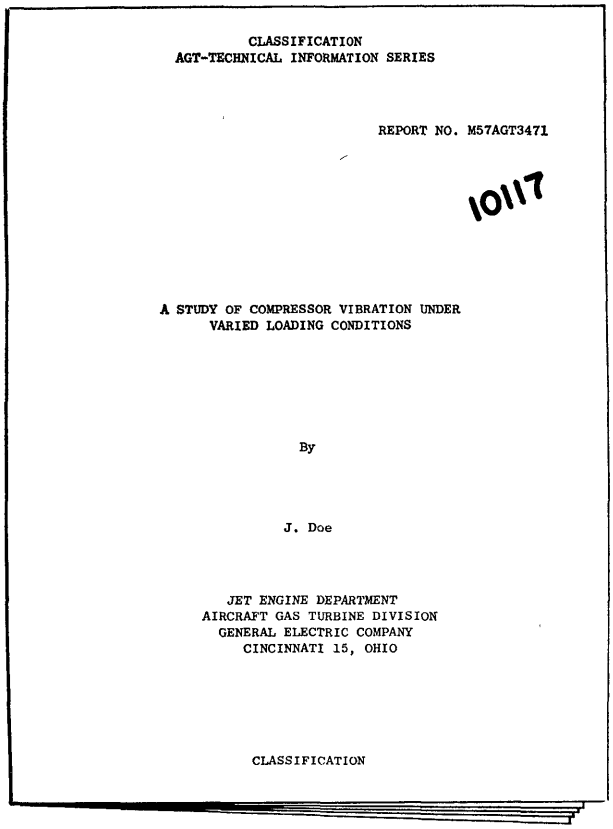
Cross coordinating, at present, is done on a strict AND system. That is, an access number must appear under each uniterm used in the search. The possibility of an AND/OR system was considered and rejected. However, the AND/OR approach is being used on a modification of this program for a different type 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.

† General Electric Co., Cincinnati, Ohio.



CLASSIFICATION
AGT-TECHNICAL INFORMATION SERIES

REPORT NO. M57AGT3471

10117

A STUDY OF COMPRESSOR VIBRATION UNDER VARIED LOADING CONDITIONS

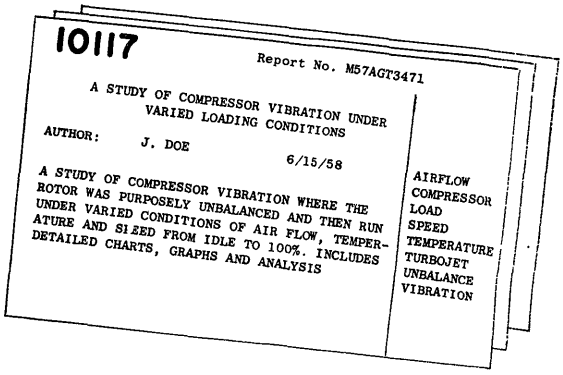
By

J. Doe

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

CLASSIFICATION

Fig. 1—Title page of AGT Report.



10117

Report No. M57AGT3471

A STUDY OF COMPRESSOR VIBRATION UNDER VARIED LOADING CONDITIONS

AUTHOR: J. DOE

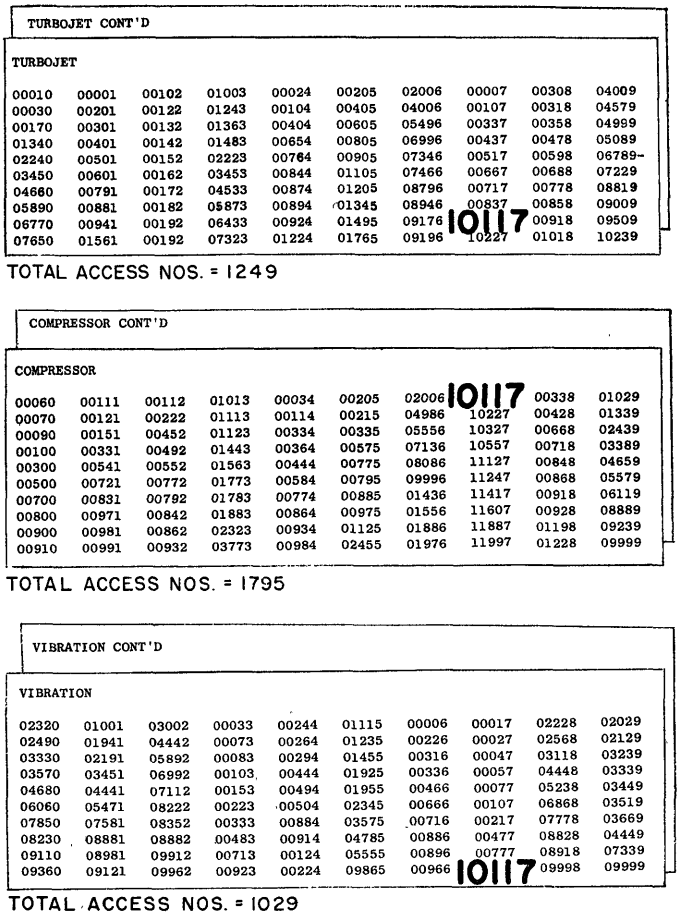
6/15/58

A STUDY OF COMPRESSOR VIBRATION WHERE THE ROTOR WAS PURPOSELY UNBALANCED AND THEN RUN UNDER VARIED CONDITIONS OF AIR FLOW, TEMPERATURE AND SPEED FROM IDLE TO 100%. INCLUDES DETAILED CHARTS, GRAPHS AND ANALYSIS

AIRFLOW
COMPRESSOR
LOAD
SPEED
TEMPERATURE
TURBOJET
UNBALANCE
VIBRATION

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



TURBOJET CONT'D

TURBOJET

00010	00001	00102	01003	00024	00205	02006	00007	00308	04009
00030	00201	00122	01243	00104	00405	04006	00107	00318	04579
00170	00301	00132	01363	00404	00605	05496	00337	00358	04999
01340	00401	00142	01483	00654	00805	06996	00437	00478	05089
02240	00501	00152	02223	00764	00905	07346	00517	00598	06789-
03450	00601	00162	03453	00844	01105	07466	00667	00688	07229
04660	00791	00172	04533	00874	01205	08796	00717	00778	08819
05890	00881	00182	05873	00894	01345	08946	00837	00858	09009
06770	00941	00192	06433	00924	01495	09176	10117	00918	09509
07650	01561	00192	07323	01224	01765	09196	10227	01018	10239

TOTAL ACCESS NOS. = 1249

COMPRESSOR CONT'D

COMPRESSOR

00060	00111	00112	01013	00034	00205	02006	10117	00338	01029
00070	00121	00222	01113	00114	00215	04986	10227	00428	01339
00090	00151	00452	01123	00334	00335	05556	10327	00668	02439
00100	00331	00492	01443	00364	00575	07136	10557	00718	03389
00300	00541	00552	01563	00444	00775	08086	11127	00848	04659
00500	00721	00772	01773	00584	00795	09996	11247	00868	05579
00700	00831	00792	01783	00774	00885	01436	11417	00918	06119
00800	00971	00842	01883	00864	00975	01556	11607	00928	08889
00900	00981	00862	02323	00934	01125	01886	11887	01198	09239
00910	00991	00932	03773	00984	02455	01976	11997	01228	09999

TOTAL ACCESS NOS. = 1795

VIBRATION CONT'D

VIBRATION

02320	01001	03002	00033	00244	01115	00006	00017	02228	02029
02490	01941	04442	00073	00264	01235	00226	00027	02568	02129
03330	02191	05892	00083	00294	01455	00316	00047	03118	03239
03570	03451	06992	00103	00444	01925	00336	00057	04448	03339
04680	04441	07112	00153	00494	01955	00466	00077	05238	03449
06060	05471	08222	00223	00504	02345	00666	00107	06868	03519
07850	07581	08352	00333	00884	03575	00716	00217	07778	03669
08230	08881	08882	00483	00914	04785	00886	00477	08828	04449
09110	08981	09912	00713	00124	05555	00896	00777	08918	07339
09360	09121	09962	00923	00224	09865	00966	10117	09998	09999

TOTAL ACCESS NOS. = 1029

Fig. 3—Uniterm wheel cards.

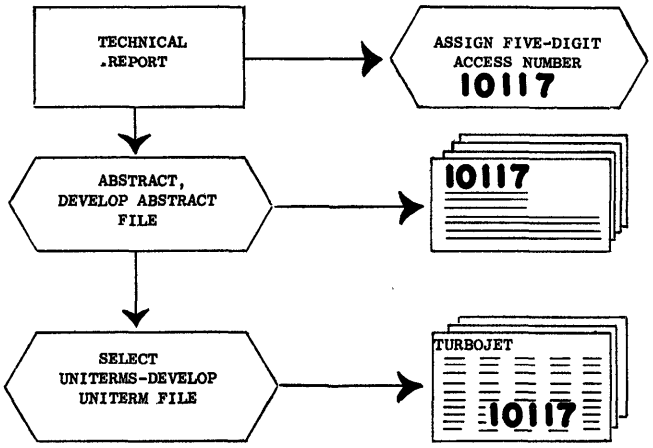


Fig. 4—Flow chart of manual system.

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

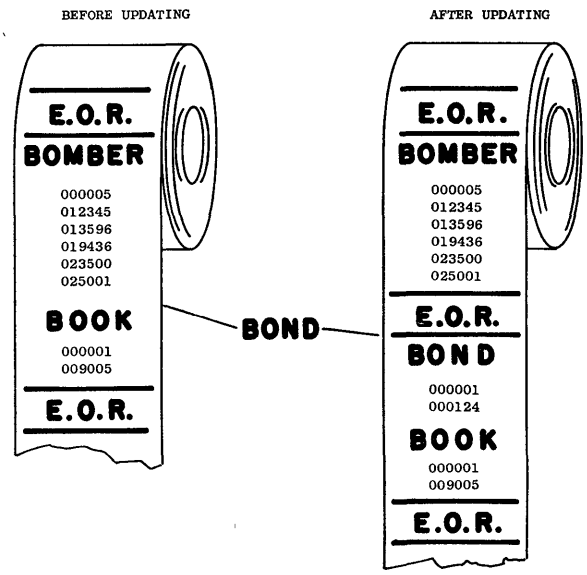


Fig. 5—Updating uniterm master tape.

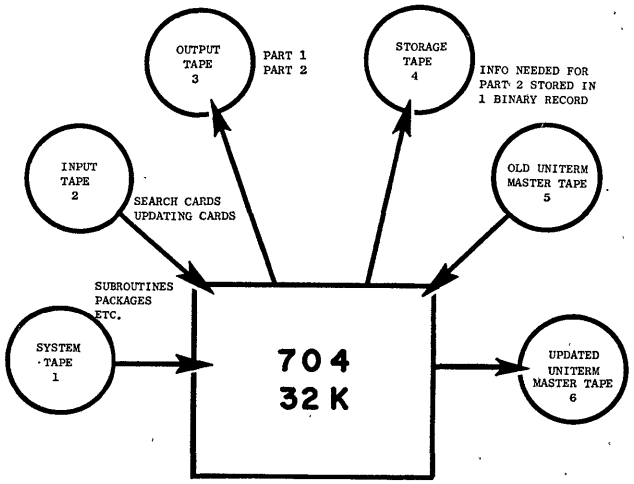


Fig. 6—Tape assignment part 1.

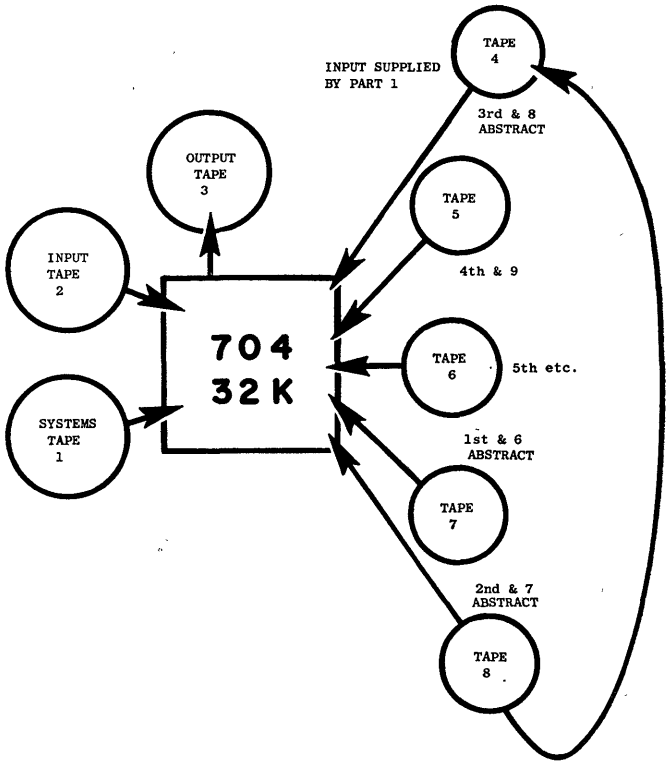


Fig. 7—Tape assignment part 2.

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 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 found 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, another abstract tape can be mounted. 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.

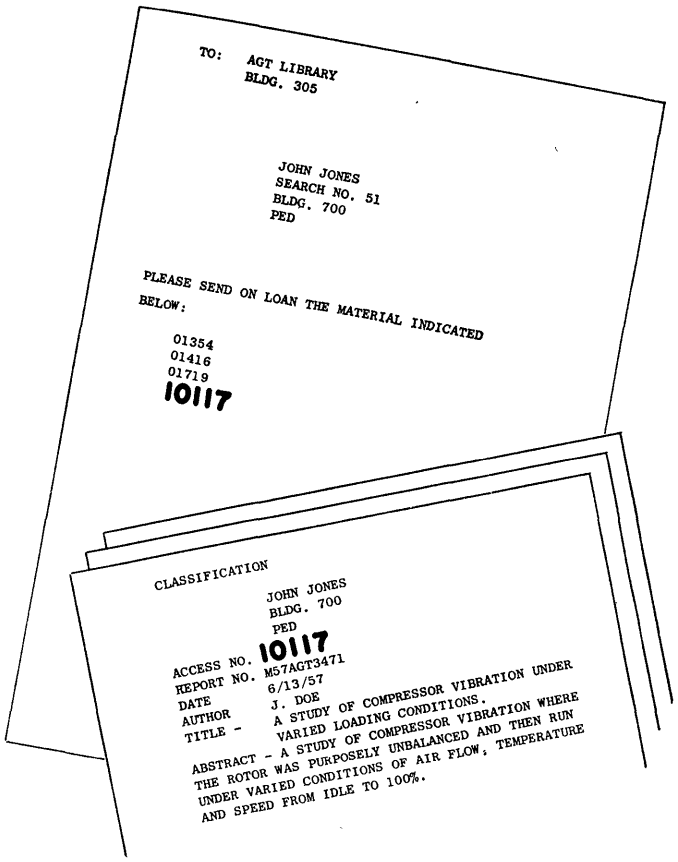


Fig. 8—704 output parts 1 and 2.

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.