XII. POSSIBLE TIME-SHARING ORGANIZATION FOR A SMART RETRIEVAL SYSTEM

Joseph J. Rocchio

1. Introduction

An important characteristic of traditional document retrieval techniques is the large amount of feedbacks which the user obtains in the course of a retrieval operation. For example, when scanning an index file, the user has an opportunity to recast his query as he correlates his original conception of the indexing system with the actual document citations observed under various index headings. If a retrieval search is made by a trained intermediary, it is again possible to modify the search prescriptions by an information exchange between user and intermediary.

In many automatic document retrieval systems the capacity for user interaction with the system is quite limited. The advent of operational computer time-sharing systems does, however, promise a much higher degree of man-machine interaction. The present note deals with some techniques needed to adapt document retrieval processes to a time-sharing environment.

2. Time-Sharing and Retrieval Systems

To establish the value of real time interaction in mechanized document retrieval, and to investigate the effects of the various types

of feedback which might be provided to the user, three basic facilities are required:

(1) a time-sharing system to effect on-line man-machine communication at a reasonable cost, such as the Compatible Time-Sharing System (CTSS)† available at M.I.T. for the IBM 7094;

(2) a flexible set of document retrieval programs which are capable of incorporating information received from the user and of providing alternative retrieval operations;

(3) a communication-oriented executive routine designed to provide the man-program interaction and to control the scheduling of the various processing options in a manner compatible with the time-sharing system.

The retrieval system described in other sections of this report (SMART) satisfies requirement (2) and, since the procedures are programmed for the IBM 7094, the system is compatible with the M.I.T. CTSS. The design of an executive routine which meets requirement (3) will then permit various tests to be made to ascertain the value of close man-machine interaction in retrieval systems.

Faced with a need for information, a user must determine a strategy to be used in the retrieval operation. The methods adopted will be influenced by the user's individual value judgments, past experience, and the nature of his needs. In view of the large variety of requirements, a

---

†The Compatible Time-Sharing System, A Programmer's Guide, The M.I.T. Computation Center, The M.I.T. Press, Cambridge, Mass. (1963).

mechanized document retrieval system should allow for a diversity of attacks so as to satisfy a large class of user needs. One important characteristic is the need for continuous feedback between the user and the system during the course of the various retrieval operations. By providing a high level of communication with the user, and enabling him to control the course of the retrieval procedures on the basis of the information he obtains, the chances of achieving satisfactory performance will be increased.

The retrieval system described in this report, operating in a time-sharing mode, has the potential of providing the kind of service mentioned above. One of the main features of the SMART system is the variety of techniques which are available for modifications of retrieval request by involving a number of statistical, semantic, and syntactic procedures. While at the present time it is not clear that all of these techniques are needed in the sense that they form an independent, mutually exclusive, set (the evaluation of these procedures is one of the main objectives of the research on the current system), it is hoped that some subset of these operations will enable effective request modification.

3. Operational Aspects

The basic retrieval procedure of the SMART system is correlation of a query vector with a set of document vectors. The result of this operation is a list, rank ordered by correlation coefficient, of documents. The user presented with this list, or some portion of it, must

estimate the adequacy of the retrieval result with respect to his needs. Since he presumably does not know the complete contents of the collection, he is operating under some degree of uncertainty. If he is not satisfied, or requires additional information to reduce the degree of uncertainty, a number of alternatives are possible.

Consider, as an example, the following situation: Suppose that as a result of the first retrieval operation the information presented to the user consists of N_i documents, of which the first n_i are listed in detail (See Fig. 1).

Correlation Level	Number of Documents Retrieved	Printed Document Citations
High	N_1	n_1 highest
Medium	N_2	n_2 samples
Low	N_3	n_3 samples

Typical Retrieval Results

Figure 1

The correlation cutoff levels in Fig. 1 are based on previous experiments and the structure of the request vector, and the numbers (n_1,n_2,n_3) of printed citations depend on N_1, N_2, and N_3 respectively.

Depending on the needs and values of the user and on his interpretation of the observed results one of the following may apply:

(1) The user is satisfied with the retrieved documents, in which case the process terminates or the complete set (N_1) of highly correlated references may be requested as output.

(2) The user identifies one or more documents within the retrieved set which he deems relevant to his request, and he would like to have the system produce those documents within the collection strongly related to the ones selected. In this case the system must find a cluster, or clusters, about elements of the selected set, or alternatively, must use in turn each of the selected documents as requests, and produce the intersection of the resulting sets of answers.

(3) The user desires to alter his request by modifying his original query on the basis of the observed citations.

(4) The user is satisfied with the basic structure of his request but would like the system automatically to expand or contract the set of retrieved documents on the basis of the observed performance.

In all cases where additional system processing is requested, the user obtains new output results and a further set of alternatives. There may, for example, be several transformations available for narrowing the scope of a retrieval prescription. In this case, if requested to provide automatic restriction, the executive routine might call in one of these transformations, say a relatively moderate contraction. If a restriction is again requested, after the original transformation has been applied, a stronger transformation might be carried out next. Another possibility is that on the basis of the prior statistical evidence, the system could decide in each case the most appropriate transformation to be used to comply with the user's chosen alternative.

In any case, by providing the ability to choose from among a sequence of retrieval operations, based on the information the user obtains while progressing through the various levels, a more effective document retrieval service can be provided for a wider class of user needs.

4. Implementation

Two basic phases are necessary for the implementation of this proposed system on a system such as the M.I.T. Compatible Time-Sharing System. First, those portions of SMART to be used in the time-sharing version must be altered to conform to CTSS conventions. In particular, this requires extensive modification of input-output orders to convert from tape operation to disk operation. Since all document processing and library maintenance programs now provided in SMART need not be included in the time-sharing versions (e.g., document and library files may be loaded off-line onto the disks), the effort required for this phase is not extensive. Second, an operational procedure for time-shared request processing must be developed, and an executive routine must be written to implement this procedure, and to provide the coordination of the various request processing subprograms which the current executive (CHIEF) provides in the SMART system.

Two sources of possible difficulty must be mentioned here. One arises in connection with the syntactical processing sections of the system (Sec. VI of this report) due to the size and time requirements of these programs. It is not clear at present how many modifications would be required to incorporate these syntactic programs into a CTSS

type system. (This would depend to some extent on the evolution of the CTSS system in the direction of allowing greater magnetic tape usage.) Furthermore, it is not determined how the response delay would be affected by incorporation into a time-sharing system. Initially, however, due to the wealth of system processes available (other than syntactic processing), this will not be a serious drawback.

The second potential difficulty arises in connection with the fact that the presently available request processing subprograms alone, together with the required data files, constitute more than a full core load. The current system is therefore run as a chain job under the FMS monitor on the regular IBM 7094. As there is no direct equivalent of a chain organization on CTSS, it will be necessary to change certain of the communication conventions which are implicit in the current system (e.g., leaving data in COMMON between individual chain links).

Two possibilities exist for implementing the kind of sequence dependent control which the CTSS system allows. One is to provide the user with a specialized language which is interpreted by the executive and results in the selection of one of the set of processing options. The other is to incorporate the effective syntax of such a language explicitly into the executive, in which case the user is given an explicit set of alternatives and is told merely to indicate his choice. This latter approach has two advantages: first, the user need have no advance knowledge in order to use the system; and, second, the implementation of the executive is simpler since explicit executive control over the branching structure required for path selection can be

maintained. Assuming that this latter approach is to be taken and that the necessary modifications can be made to the selected programs of SMART, the mode of operation may be outlined as follows.

The user, at a remote CTSS console creates a request file on the disk by use of the built-in CTSS commands (INPUT,EDIT). When the user is satisfied that the request(s) are error-free the FILE command is issued, to create a disk file consisting of a sequence of card images corresponding to the entered text.

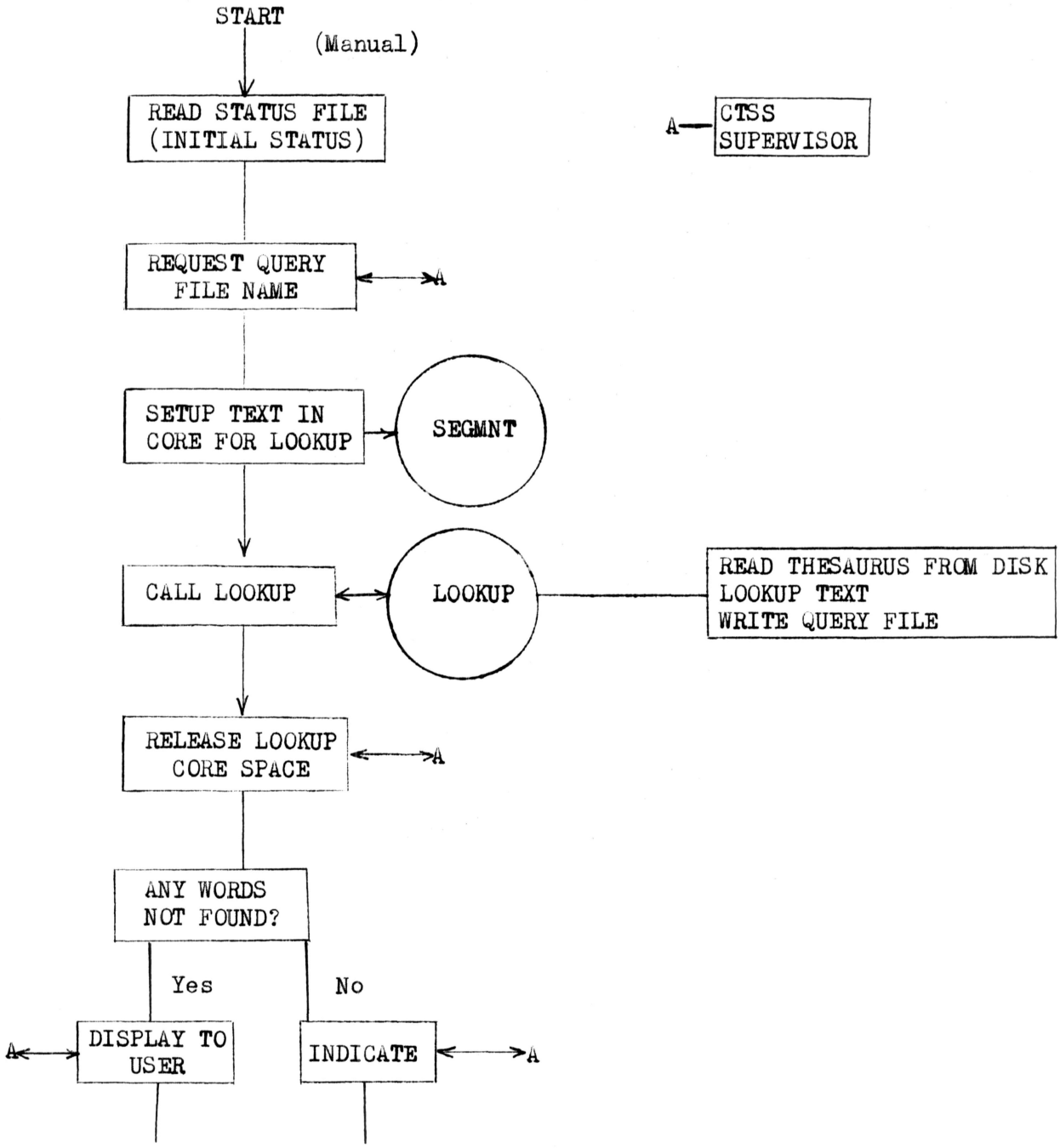
At this point the retrieval process may be initiated. The user, again via the CTSS commands, causes initial configuration programs to be loaded into core. The executive requests from the user the name of the query file and then calls the input routine (SEGMNT) to read this file into core. After the query has been read, the lookup program is entered. This program reads in the thesaurus from disks and generates a concept vector for the query. This vector is written out as a special file, and the status information concerning the lookup is left in core. Upon return to the executive the core space required by lookup and thesaurus is released (to improve response time), and the status of the lookup is tested. If any words were not found in the thesaurus, these are displayed to the user and he is given the option of proceeding or modifying his request. If all words were found this fact is reported and the supervisor automatically proceeds to request document correlation.

This type of proceeding is made possible by providing a special status file maintained by the supervisor on disk. At the start of each processing step, the first task of the executive is to read this status

file, and to determine from it the function of the current retrieval phase. Thus, at the end of the lookup phase, the executive and correlation programs are loaded. The executive then reads the status file and sets up the document and query vectors in core. After the correlation programs generate a query-document correlation matrix on a disk file, the executive can release the excess core space and present the results to the user. Again, the user is given a set of alternatives applicable at this point. When a choice is made by the user, the executive updates the status file and chains into the LOADGO CTSS command with the appropriate parameter to load the set of programs required for the next processing option. Thus, at the termination of each processing phase, the executive sets up the data for the next phase chosen by the user, and then causes the appropriate programs to be read in from the SMART program files on disk.

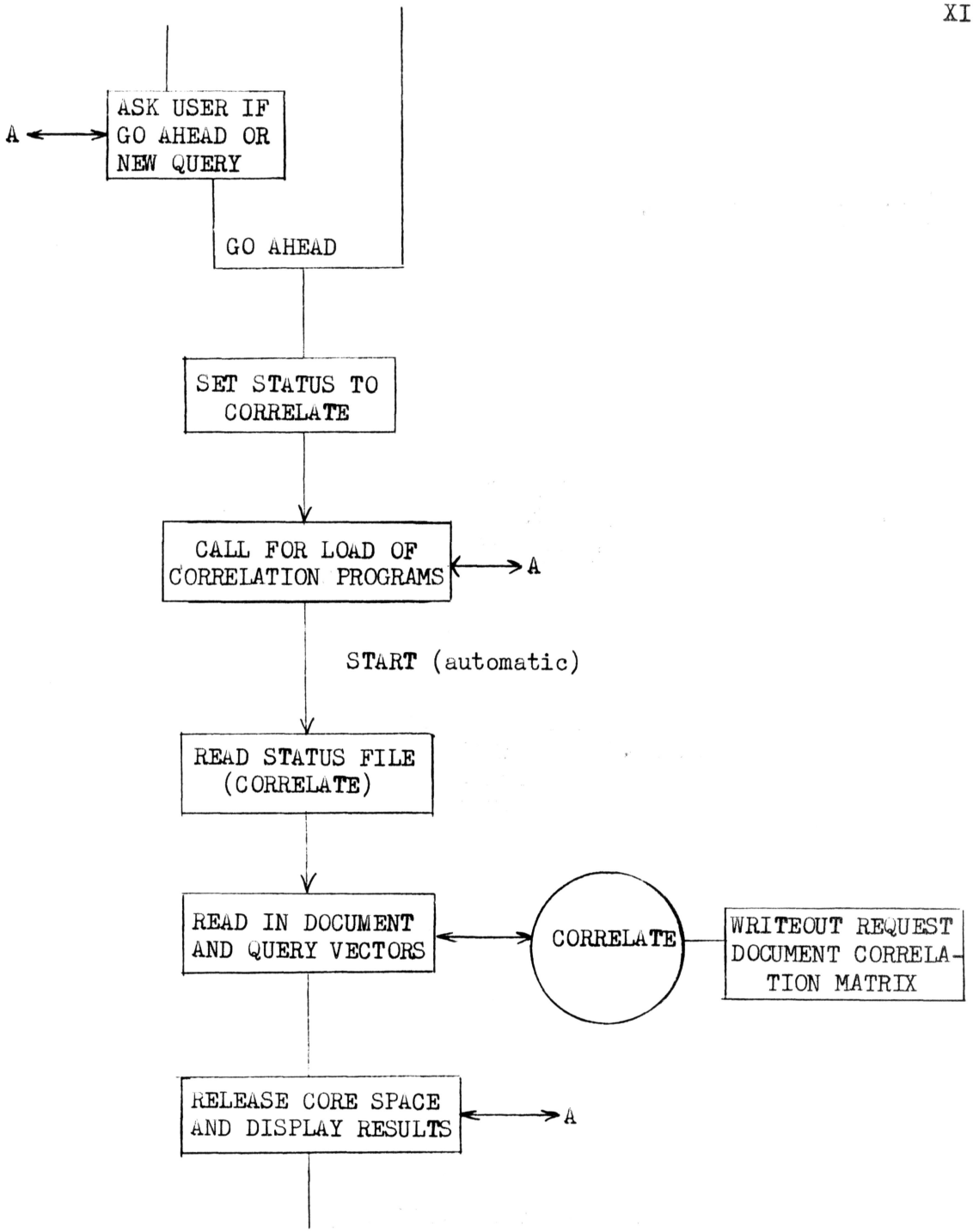
A rough block diagram of this procedure is given in Flowchart 1. Additional effort is required in designing detailed specifications, and in studying the feasibility of such a system organization.

A hypothetical example of retrieval sessions at a CTSS console is given in Appendex A.

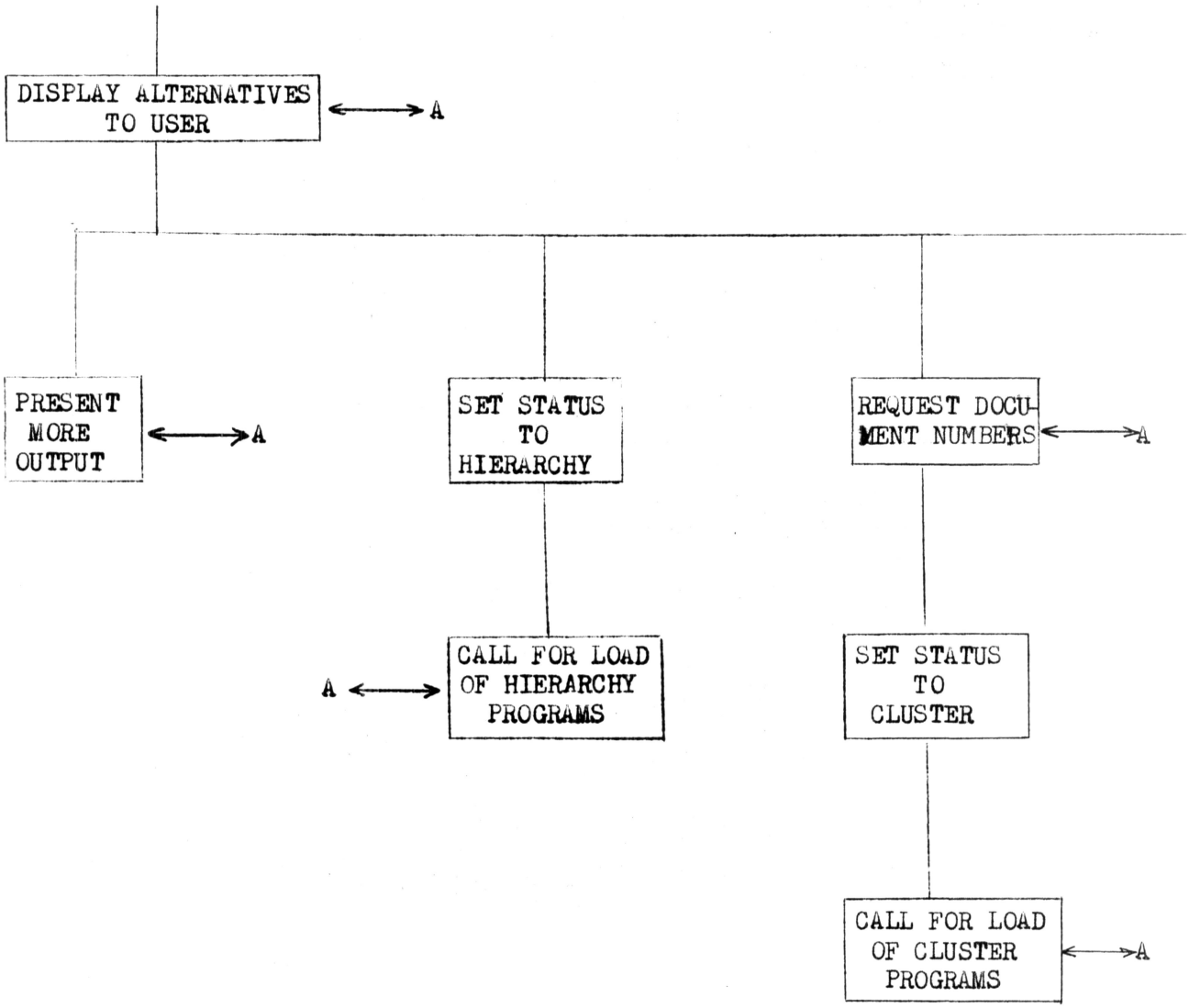


Possible Time-Sharing Organizations for SMART

Flowchart 1



Flowchart 1 (continued)



Flowchart 1 (continued)

APPENDIX A

As a further illustration of the type of retrieval service the proposed time-sharing system might provide, an example is given below of a hypothetical session of a user at a CTSS console. The retrieval results shown are based on initial experiments with the present SMART system; the document file used for these experiments consisted of fifty document abstracts taken from the collection contained in the IRE PGEC Transactions.

In the description below user input is capitalized and indented. Computer output from CTSS or the SMART system appears in lower case. Explanatory notes are enclosed in parentheses.

(After logging in, the user is ready to type in his query. At this point the conversation is between the user and CTSS.)

... INPUT

00010 ... *TEXT ANALOG-TO-DIGITAL

00020 ... ANALOG TO DIGITAL CONVERSION. DEVICES AND
TECHNIQUES FOR THE CODING

00030 ... OF ANALOG INFORMATION.

00040 ... π (puts CTSS in manual input mode)

(The first line identifies the query to the SMART system. The user now creates a disk file named QUERY by the following CTSS command.)

... FILE QUERY DATA

(At this point retrieval may be initiated; the user causes loading of the initial program segment.)

... LOAD SMART
... START
(Responses to the user from this point are from the SMART system.)

enter the name of the query file

... QUERY

all words found by lookup

retrieval results by query-document correlation:
query: analog-digital
answers:
correlation .66, 2 documents
296 a high-speed analog to digital converter
132 logical synthesis of some high-speed
digital comparators

correlation .55, 6 documents, one of which is:
229 matrix programming of electronic analog computers

correlation .44, 6 documents, one of which is:
62 on programming of arithmetic operations

options:
to alter query type *end* and use ctss editing commands.

For display of all answers with
correlation above N type *print N*

for expansion or contraction by hierarchy
type *hier expand* or *hier contract*

for document clustering type *clust n_1,n_2*
with as many document numbers as desired
separated by commas.

to terminate retrieval type *end*.

... PRINT .5

additional answers with correlation above .5
322 uses of compiler programs to solve power problems
289 stable high-speed digital-to-analog conversion for storage tube deflection
263 machine language in digital computer design
255 three levels of linguistic analysis in machine translation

options:

same as those above.

... CLUST 296, 289

document cluster of 296, 289:
cluster contains documents 296, 289.

options:

same as those above

... END