SUMMARY

A detailed description of the SMART automatic document retrieval system which became operational on the IBM 7094 during the summer of 1964 was included in a previous report in this series (Information Storage and Retrieval, Report No. ISR-7, dated June 1964). The present report, like its predecessor, is also again devoted to a description of work conducted in connection with the SMART system over the past several months. Two aspects of the system are stressed in particular: its potential as an operating, fully automatic, text processing and retrieval system; and its use as a testing device to evaluate the effectiveness of a large variety of automatic content analysis procedures.

Sections I and II by M. Lesk are concerned with a description of SMART as an operating retrieval system. In Sec. I, the systems operations are described in detail; a typical search request and a typical set of analysis methods are used for purposes of illustration. Actual computer-produced output is reproduced to show how search requests and incoming documents are introduced into the computer; the indexing products which are automatically generated by several different analysis procedures are then exhibited, followed by the correlations obtained as a result of the comparison between analyzed documents and search requests. Output is shown in several different formats, and the analysis methods are compared by an automatic evaluation process which matches manually generated relevance judgments with the computer output actually obtained.

Section II contains detailed operating instructions which should make it possible to prepare data and program decks, and to run the system at installations with the compatible computer and compatible monitor systems. Suggestions are also included for adapting the available programs to operating environments which may differ somewhat from those prevailing at Harvard.

Sections III, IV, and X are devoted to the general problem of retrieval systems evaluation. In Sec. III by J. Rocchio, a set of new functions is introduced designed to measure the effectiveness of retrieval performance. These new measures are not dependent on a possibly artificial distinction between retrieved and nonretrieved information, as are the standard recall and precision (relevance) measures. Instead, they are based on the rank orders of the relevant documents, when the document collection is arranged in decreasing correlation order with the search requests.

Section IV by G. Salton contains the initial results of the evaluation process obtained with the SMART system over the past several months. The data reported cover approximately 20 search requests used with a collection of 405 document abstracts, and a set of about 15 different analysis procedures. It is found that methods using complete abstract texts are superior to those which confine the analysis to the titles only. Use of a synonym dictionary to normalize the vocabulary gives much better results than use of the original words occurring in documents and search requests.

Finally, the phrase procedures based on the use of combinations of terms, rather than individual terms alone, seem more powerful than most other available procedures. The effects of the iterative operating process which makes it possible to analyze the same search request in several ways and to combine results are also outlined.

Section X by J. Rocchio and M. Engel describes the evaluation process in more detail, and presents in tabular form the retrieval results obtained for each search request and each processing method used.

The phrase-matching procedures incorporated into the SMART system are believed to be particularly powerful, and have not previously been used for retrieval purposes. In Sec. V by G. Salton, the phrase generating process is therefore discussed in more detail, and it is shown how the phrase-matching methods make it possible effectively to assign the same content identifications to hundreds of semantically equivalent but syntactically quite different constructions. The use of phrase matching as a partial replacement for a sentence kernelizing routine is also discussed.

The "criterion phrases," consisting of combinations of syntactically related terms, are stored in the computer in a special predetermined format. This format makes it possible to identify each phrase component by sets of semantic and syntactic markers, and by the syntactic relationships which arise between the various components. The format specifications used for the

construction of the criterion phrase dictionary are described in detail in Sec. VI by A. Lemmon.

The SMART system requires for its operations a set of five principal dictionaries: alphabetic stem and suffix dictionaries, a concept hierarchy in tree form, and statistical as well as syntactic (criterion) phrase dictionaries. These dictionaries are originally set up by hand. Section VII by C. Harris describes some of the problems which arise in the original construction and in the updating of these dictionaries.

Section VIII by T. Evslin and IX by M. Cane are devoted to various extensions of the SMART retrieval system which are presently being actively considered. The programs operating at the present time suffer from various size restrictions which limit the usefulness of the system. In particular, since the dictionary lookup routines operate in core, and require simultaneous storage of the complete thesaurus (alphabetic dictionary) and of one complete document at any given time, the size of the thesaurus as well as the permissible document lengths are strictly limited. Limitations are also imposed on the permissible size of the complete document file. These size limitations are presently being removed. Flowcharts and descriptions for this "extended" SMART system are included in Sec. VIII.

It may be expected that in the foreseeable future most practical information systems will operate under some type of user control. A typical

organization might consist of a set of consoles which could simultaneously accommodate a number of users communicating with a central retrieval device. The operations of the SMART system can be adapted to such an organization, as is described in more detail in Sec. IX. In particular, a supervisory system is introduced in that section, as well as a set of input-output type-in and print-out statements, which are designed to adapt existing operations to the M.I.T. compatible time-sharing system (CTSS). Programming efforts in this direction are presently under way.