SUMMARY

The present report, number ISR-7, is devoted almost entirely to a detailed description of the SMART experimental document retrieval system. This system has been programmed at the Computation Laboratory of Harvard University over the past two years and is now in operation on the IBM 7094. The report is written largely from a programmer's viewpoint and a sufficient number of operating details are included to enable the reader to use the programs in his own installation. It may be expected, however, that additional instructions will be needed if large-scale modifications of the present routines are to be carried out successfully.

Sections I and II by G. Salton and M. Lesk, respectively, contain an overall description of the system. Section I concentrates on the system philosophy, the reasons behind the chosen organization, and the expected uses and advantages of the system. Section II, on the other hand, contains a general program description, including the organization of the supervisory system (CHIEF) and of the subroutine calling sequences. Operating characteristics of the system are also included in Sec. II.

Sections III, IV, and V by C. Harris, M. Cane and G. Shapiro, respectively, deal with the construction, use, and updating of the word dictionaries and of the concept hierarchy incorporated into the system.

Section III outlines the main features pertaining to the organization of the dictionaries. Section IV describes the dictionary lookup procedures, including the programs which separate text words into stems and suffixes, the replacement of these stems and suffixes by appropriate syntactic and semantic codes, and the updating methods provided for the maintenance of stem and suffix dictionaries. Section V is intended to do for the concept hierarchy what Sec. IV does for the word dictionary. Specifically, a number of list-processing methods are outlined which may be used to traverse the tree-like organization of the concept hierarchy by going "up," "down," or "across" in the tree. Also included in Sec. V is a description of the elaborate hierarchy up-date options permitting deletion and addition of items and proper maintenance of the link addresses.

Sections VI, VII, and VIII by A. Lemmon, E. H. Sussenguth, and T. Evslin and T. Lewis, respectively, deal with the syntactic procedures. Section VI contains a description of the editing features needed to prepare the input for the syntactic analysis programs. The methods used to compare syntactically analyzed text excerpts with the criterion tree dictionary are also outlined in Sec. VI. The actual graph matching algorithm used for the sentence comparisons is briefly described in Sec. VII, and the construction and updating of the criterion tree file are covered in Sec. VIII.

The statistical procedures incorporated into the SMART system are described in Secs. IX and X by M. Lesk and T. Evslin, and M. Lesk, respectively. Section IX covers a set of programs designed to recognize certain concept sets known as "statistical phrases;" the components of such phrases are not related syntactically, but are defined by their co-occurrence characteristics within the sentences of the various documents. Additional statistical processing methods, including the generation of term-term and document-document similarity coefficients, and the production of term and document clusters are described in Sec. X.

Section XI, the last in this series, is concerned with the description of a variety of housekeeping routines; some of these are internal to the system, and are as such inaccessible to the user. Specifically included in Sec. XI is a system for maximizing the available storage space by clearing those areas in memory which are not needed in any given run. The construction of the "vacuous" dictionary which may be used to assign dummy concept numbers to the text words is also described in Sec. XI.

The last three sections included in this report, No. XII by J. J. Rocchio, and Nos. XIII and XIV by A. R. LeSchack do not properly cover any part of the existing SMART systems, but deal with extensions to the system and with additional research of a statistical nature. Section XII describes some of the problems which must be solved before a retrieval system such as SMART can be adapted to a time-sharing organization in which several users are tied to the same central equipment by means of special input-output units and communications channels. The "cosine" similarity measure available in SMART to aid in the statistical processing is examined in Sec. XIII, and it is shown that it is compatible with the classical product-moment correlation coefficient in all cases likely to arise in practice. Finally, a new clustering procedure, using matrix eigenvalue analysis and capable of grouping similar sets of objects, such as item properties or documents, is introduced in Sec. XIV, and the relations of this new method to other standard clustering algorithms are shown.