Summary

The present report is the fourteenth in a series describing research in automatic information storage and retrieval conducted by the Department of Computer Science at Cornell University with the assistance of the Division of Engineering and Applied Physics at Harvard University. The report covering work carried out on the SMART project for approximately one year (summer 1967 to summer 1968) is separated into four main parts: Smart systems design (sections I and II), analysis and search experiments (sections III to VI), user feedback procedures (sections VII to XIII), and descriptions of text editing programs (sections XIV and XV).

During the past year, the main effort has again been devoted to experiments designed to determine the effectiveness of the automatic search and retrieval techniques incorporated into the SMART system. A comparison has thus been made between the fully automatic text processing methods used by SMART and the partly manual methods in use by the Medlars system operating at the National Library of Medicine (section VI). Furthermore, increasing attention has also been given to interactive search procedures based on user feedback information supplied during the search process to refine the query formulations (sections VII to XIII). Finally the evaluation methodology incorporated into the SMART system has been examined, and evidence is supplied indicating that the evaluation parameters in use may be largely invariant with alterations in the query-document relevance judgments (section III).

The conversion of the SMART text processing programs from the present 7094 implementation to operations using an IBM 360/65 system is

continuing (section II). At the same time, steps have been taken to implement a time-sharing version of the SMART system using input-output consoles to introduce data and search requests. The time sharing implementation is expected to accommodate both on-demand searches carried out in real time while the customer is waiting as well as background batch processing work performed when no user interaction takes place (section I).

Section I by M. E. Lesk covers a proposed design for automatic on-line information retrieval operations. The proposed system uses disk packs and data cells to store the document files, and dedicated consoles for file interrogation. A supervisiory system chooses among the jobs to be processed by checking the job priority as well as the accessibility of the required files. Procedures are outlined for text input; request and text look-up; thesaurus, phrase, and hierarchy processing for English as well as foreign language data; concept vector formation; document clustering; feedback procedures; dictionary displays; and other auxiliary functions.

Section II by D. Williamson summarizes the design of the batch processing implementation of the SMART programs on the IBM 360/65. The program implementation stresses operating speed by providing for the parallel execution of several file searches; flexibility of use is obtained by making it possible at any given time to choose among a large number of alternative processing methods.

Sections III to VI deal with the design and evaluation of the basic SMART analysis and search procedures. The accuracy and trustworthiness of the evaluation output included in this, as well as in previous reports in this series is examined in section III by M. E. Lesk and G. Salton. Using a collection of 1200 documents in the area of documentation as a test collection, it is shown that although the agreement among relevance assessments supplied by different user populations is only about thirty percent, the recall-precision graphs produced by the evaluation process do not vary with alterations in the relevance judgments. Thus, the conclusions drawn from the SMART experiments are apparently applicable to many different user classes.

Under normal circumstances, the automatic text analysis methods incorporated into SMART do not include complete disambiguation procedures for every text item. An attempt is made in section IV by M. Coyaud to construct disambiguation rules leading to a unique identification of all terms taken from a corpus in ophthalmology. The disambiguation rules are based on contextual criteria furnished by the grammatical information attached to the text words. The efficacy of the disambiguation rules remains to be compared with the effectiveness of the text analysis methods normally used in the SMART system.

The search methods to be used in a real-time retrieval system (where fast response times become mandatory) may be expected to be based on a clustered document collection where only certain document clusters are compared against each search request. Most clustering procedures are, however, expensive to carry out, since the time varies with the square of the number of items to be clustered. A fast clustering algorithm of order n log n is examined in section V by R. T. Dattola, and its properties are outlined. The algorithm is presently being implemented as a part of the SMART retrieval procedures.

In section VI by G. Salton and D. K. Williamson a comparison is performed between the effectiveness of the fully automatic text processing methods used by SMART, and the conventional procedures based on manual document and query analysis used by the Medlars system at the National Library of Medicine. The test collection consists of 18 queries and 273 documents in biomedicine. It is found that the recall figures are comparable for the two systems, while the adjusted precision is somewhat smaller for SMART than for Medlars. It is expected that the test can be extended to larger query and document samples in the near future.

Interactive retrieval methods based on feedback information supplied by system users are examined in sections VII to XIII. Sections VII and VIII by G. Salton and E. Ide, respectively, cover recent experiments using the relevance feedback methods previously introduced in reports ISR-11, and ISR-12. Here the user supplies relevance judgments for documents previously retrieved by the system; these judgments are then used automatically to update the queries so as to render them more similar to items previously identified as relevant, and less similar to the nonrelevant items. A variety of feedback methods are treated, and the effectiveness of each is discussed.

Section IX by M. E. Lesk and G. Salton deals with interactive procedures where the queries are not updated automatically by the system, but the user is forced to undertake the query alteration himself based on information displays provided during the search process. The interaction can take place either prior to an actual file search using a display of term frequencies, or thesaurus excerpts, or source documents; alternatively, post-search interaction can be based on a display of titles or texts of previously retrieved documents. The usefulness of various alternative query alteration methods is examined and evaluated as a function of search effectiveness, user effort, and retrieval cost.

In the interactive retrieval procedures treated up to now, only the user queries are actually altered, following the user-system interaction. The document space remains invariant throughout. It is, however, possible to carry out searches by using both query and document space alterations. Two alternative procedures including document as well as query modification are examined in section X by M. C. Davis, M. D. Linsky, and M. V. Zelkowitz, and section XI by T. L. Brauen, R. C. Holt, and T. R. Wilcox, respectively. The process described in section X performs a modification of the complete document space by computing a discrimination factor for each concept, which is then used to control the concept alteration. In section XI, on the other hand, only those documents previously judged relevant by the user are actually altered. A brief evaluation of the process is included in each section.

It is often observed that users of a retrieval system have difficulties in generating useful query formulations. In section XII by A. Borodin, L. Kerr, and F. Lewis, an automatic query splitting method is described in which the users' queries are automatically split into two or more subsidiary queries if it is determined that relevant items previously identified by the users actually belong to different document groups. Each subsidiary query is then processed separately in an attempt to retrieve different document sets. The effectiveness of the feedback method using split queries is compared with the standard feedback methodology using single queries only.

Section XIII by R. G. Crawford and H. Z. Melzer introduces an interactive process where the search is based on the identification of a "source" document known by the user to be relevant to his information need, instead of an original query formulation. It is concluded that users who know of a document definitely claimed to be relevant should produce this document to replace the original query statement.

The last two sections of the report numbered XIV and XV by J. Bean and E. R. Quinones, respectively, describe automatic text editing programs used for document and query file preparation. The editing procedures of section XIV are based on the processing of complete card images on tape in either blocked or unblocked forms, using serial numbers or references, while the procedures of section XV perform context editing of individual character strings within a card image. In each case, the editing commands are specified, and timing estimates are given for the various procedures.

Readers interested in additional material dealing with the automatic SMART analysis and search system may wish to consult the text titled "Automatic Information Organization and Retrieval" recently published in McGraw Hill's Computer Science Series, as well as reports ISR-11, ISR-12, and ISR-13 previously published in the present series. These earlier reports include sections on systems design (ISR-11, I-IV); evaluation (ISR-11,V; ISR-12, I-III; ISR-13, I-X); feedback (ISR-11, VI; ISR-12, VIII-XII); and cluster searches (ISR-11, VII, IX; ISR-12, IV-VII). The reports are available from the Clearinghouse for Federal, Scientific, and Technical Information in Springfield, Virginia.

G. Salton