Summary

The present report is the sixteenth 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 by the SMART project for approximately one year (summer 1968 to summer 1969) is separated into four main parts: SMART system design (Sections I to III), text processing and analysis experiments (Sections IV to VIII), user feedback procedures (Sections IX to XII), and document and query clustering methods (Sections XIII to XV).

Most recipients of SMART project reports will experience a gap in the series of scientific reports received to date. Report ISR-15, consisting of a master's thesis by Eleanor Ide entitled "Relevance Feedback in an Automatic Document Retrieval System" was prepared for limited distribution during the spring of 1969. Report ISR-15 is available from the Clearinghouse for Federal, Scientific and Technical Information in Springfield, Virginia 22151, under order number PB 184-246. The forthcoming report ISR-17 is similarly to be issued for limited distribution only in the fall of 1969.

During the past year, a complete SMART text processing and analysis system has been implemented on the IBM 360 model 65 at Cornell University. The standard batch processing mode of operations is eventually to be replaced by an on-line processing system using console devices for input and output. In addition to the 360 implementation, a version of the SMART system is operating on IBM 7094 equipment in several places, and several experimental SMART procedures, including notably the user "relevance feedback" methods have been incorporated into a number of operating retrieval systems both in

the United States and abroad.

Of particular interest in the present volume may be the description of the 360/65 implementation of the SMART system presently available at Cornell (Section I). In addition, the automatic text analysis methods have been extended for the first time to documents and queries not written in English. A comparison is made in Section IV between English and German document processing using identical language analysis methods with a multi-lingual thesaurus. A variety of novel interactive search procedures, based on user feedback information, and leading to alterations in both the document and the query representations are described in Section IX. Finally, Sections XIII and XIV cover new, efficient clustering methods designed to group the documents in a collection into affinity classes. Evaluation results are given comparing resulting partial cluster searches with full searches of the complete document collections.

Sections I to III cover existing or proposed design features of the SMART document retrieval system. Section I by Donna Williamson is a description of the IBM 360 implementation of the SMART system, including text processing, language analysis, query-document matching, file organization and clustering, user feedback procedures, and retrieval evaluation methods. Sample output is shown to illustrate present system capabilities.

Section II by D. M. Murray, describes a variety of scatter storage systems useful for dictionary storage and look-up procedures. The normal ordered dictionary storage systems (which arrange English text words in alphabetical order) produce fast searching but slow dictionary updating; chained storage systems, on the other hand, are easy to update but slow to search. Scatter storage methods which transform input words into so-called

"hash addresses" provide fast access and considerable storage economies. Several possible implementations of scatter storage systems are described and their characteristics in a retrieval environment are outlined.

A new single parameter evaluation measure is introduced in Section III by J. Joiner and L. Werner. Criteria are first listed to measure the usefulness of retrieval evaluation parameters. A new, single parameter probability measure is then derived which includes the effects of recall, precision and generality, and consists of the probability, under the hypergeometric distribution, that for a given number of items retrieved the precision could be strictly less than that actually attained. Sample calculations are shown for several possible retrieval runs.

Several different text processing and content analysis experiments are treated in part 2 of this report, consisting of Sections IV to VIII. Section IV by G. Salton describes an extension of the standard automatic analysis procedures incorporated into the SMART system to documents not written in English. Specifically, a multi-lingual English-German thesaurus is used to compare the performance of English and German queries using first a collection of English document abstracts, and then a collection of German documents. The change in the query language appears to produce no substantial deterioration in retrieval effectiveness.

Section V by S. F. Weiss deals with the use of automatic syntactic analysis methods for purposes of content analysis. Various methods are described for the automatic generation of syntactic phrases, and these procedures are then tested in a retrieval environment. Small improvements in retrieval effectiveness are obtained when the use of syntactic phrases is compared with the use of statistically derived phrases based on word cooc-

currences in queries and documents.

Section VI by S. F. Weiss describes simple "template analysis" procedures to replace the more elaborate syntactic analysis methods often used for automatic text processing. Specifically, an attempt is made to identify one or more "templates", that is, previously known strings of common words, in a given text; the occurrence of a template in a given context then leads to a set of action routines designed to carry out the input translation. Various template types are examined, as well as the corresponding action routines. Procedures are then given to carry out the template matching, and their use is described in recognizing date phrases, journal phrases, and author phrases occurring in standard information requests.

Previous retrieval evaluation results have shown that a word stem analysis method is most useful to obtain high precision, whereas the use of a synonym dictionary or thesaurus leads to high recall. A content analysis method is examined in Section VII by B. Faith and J. Jensen in which thesaurus and word stem analyses are combined to produce hybrid document and query identifications. Evaluation results obtained by using 200 documents and 42 queries in aerodynamics show that slightly better normalized evaluation measures are generated with the hybrid vectors than with the standard vectors at the expense of larger storage requirements.

Section VIII by J. W. McNeill and C. S. Wetherell deals with the use of bibliographic data for document and query analysis purposes, continuing thereby an earlier investigation described in Section XI of report ISR-12. Specifically, author and publication place identifiers are added to the normal document identifications; the queries are similarly adjusted by taking bibliographic information from a set of previously identified relevant docu-

ments. An evaluation is then performed using 273 documents and 18 queries in medicine, and the effectiveness of the elongated vectors using bibliographic information is compared with the effectiveness of standard vectors including the normal content identifiers only.

User feedback procedures are treated in part 3 of this report, including Sections IX to XII. In Section IX by J. S. Brown and P. D. Reilly several selective feedback methods are described in which certain particularly significant document concepts — those which can discriminate between relevant and nonrelevant documents — are used to alter the user queries during the search negotiation process. A different weight is used for positively significant and for negatively significant concepts, and the process is evaluated using a collection of 200 documents and 42 queries in aerodynamics.

Several novel feedback evaluation methods are described in Section X by C. Cirillo, Y. K. Chang, and J. Razon. These procedures, originally outlined by E. Ide in report ISR-15, are motivated by the fact that it is necessary in a feedback evaluation to distinguish improvements in the ranks of previously retrieved relevant documents from improvements in the ranks of items not previously seen. Three different evaluation methods are covered, called respectively, the modified freezing process, the residual collection method and the test and control evaluation. Each method is examined using document and query collections in aerodynamics.

A variety of special relevance feedback strategies are described in Section XI by E. Ide and G. Salton, including in particular certain selective negative feedback methods designed to improve the feedback performance obtained from nonrelevant documents; cluster feedback, query splitting, and query clustering procedures designed to identify separated

groups of relevant documents; and document space modification techniques which alter not only the query identifications but also the identifications of certain relevant or nonrelevant documents previously identified by the user.

Finally, in Section XII by T. Leventhal and R. Miller, the query splitting process, earlier examined in report ISR-14 (Sections XII and XIII) is again taken up in an environment in which separated groups of relevant documents are used directly to search a collection, replacing a standard query splitting operation. The search results obtained by using these separated document groups as queries are then compared against conventional relevance feedback methods.

The last part of this report, consisting of Sections XIII to XV covers query and document clustering methods which may produce rapid searches of the document collections. In Section XIII, R. Dattola examines a fast clustering process which groups n items into m classes in approximately n•m operations. The algorithm is described in detail, and its effectiveness is evaluated using document collections in aerodynamics and documentation. Cluster search evaluation methods are also covered, and suggestions are made for the use of the "correlation percentage" as part of a cluster evaluation process. It is found that the effectiveness of the partial clustering process compares favorably at low recall with that of a full search method.

An even simpler one-pass clustering process is examined in Section XIV by S. Rieber and V. P. Marathé, requiring only n operations to group n items. The method is outlined and used to process a collection of 82 documents and 35 queries in documentation. Evaluation results are appended.

The last section, number XV by S. Worona, deals with a query clus-

tering procedure, originally proposed by V. Lesser in Section VII of report ISR-11. Here queries are grouped, instead of documents, and documents are then chosen for query comparison according as they correlate highly with the centers of certain query clusters, or with certain queries contained in the query clusters. The process is evaluated using a collection of 424 documents and 155 queries in aerodynamics, and comparisons are made with full searches, and with partial searches resulting from the normal document clustering procedures. The query clustering method appears to operate as effectively as the previously used standard procedures.

Additional automatic content analysis and search procedures used with the SMART system are described in several previous reports in this series, including notably reports ISR-11 to ISR-15 published between 1966 and 1969. These reports are all available from the Clearinghouse in Springfield, Virginia under order numbers PB 173-196, PB 176-536, PB 177-812, PB 180-931, and PB 184-246, respectively.

G. Salton