DEPARTMENT OF COMPUTER SCIENCE

CORNELL UNIVERSITY

INFORMATION STORAGE AND RETRIEVAL

Scientific Report No. ISR-11

to

The National Science Foundation

Ithaca, New York
June 1966

Gerard Salton
Project Director

Department of Computer Science
Cornell University
Ithaca, New York 14850

Scientific Report No. ISR-11

INFORMATION STORAGE AND RETRIEVAL

to

The National Science Foundation

Ithaca, New York
June 1966

Gerard Salton
Project Director

Staff of the Department of Computer Science

Cornell University

Richard W. Conway
Margaret Dodd
Patrick C. Fischer
Juris Hartmanis
Michael Keen
Joann Newman
Christopher Pottle
Gerard Salton
Sidney Saltzman
Robert J. Walker

Project Staff at the Aiken Computation Laboratory

Harvard University

Jeffrey Bean
Claudine Harris
Guy Hochgesang
Michael Lesk

TABLE OF CONTENTS

TABLE OF CONTENTS (continued)

Page

SECTION II (continued)

TABLE OF CONTENTS (continued)

Page

SECTION II (continued)

SECTION III

HOCHGESAŃG, G. T.: "SØCCER - A Concordance Program"

TABLE OF CONTENTS (continued)

Page

SECTION III (continued)

SECTION IV

SALTON, G., and LESK, M. E.: "Information Analysis and Dictionary Construction"

TABLE OF CONTENTS (continued)

Page

SECTION IV (continued)

SECTION V

TABLE OF CONTENTS (continued)

Page

SECTION V (continued)

SECTION VI

TABLE OF CONTENTS (continued)

Page

SECTION VII

LESSER, V. R.: "A Modified Two-Level Search Algorithm Using Request Clustering"

TABLE OF CONTENTS (continued)

Page

SECTION VIII

BLOMGREN, G., GOODMAN, A., and
KELLY, L.: "An Experimental Investigation of Automatic
Hierarchy Generation"

SECTION IX

BROFFIT, J. D., MORGAN, H. L., and
SODEN, J. V.: " On Some Clustering Techniques for Information
Retrieval"

TABLE OF CONTENTS (continued)

Page

SECTION X

LESK, M. E.: "Design Considerations for Time Shared Automatic Documentation Centers"

Summary

The present report is the eleventh in a series covering research in automatic storage and retrieval conducted initially at the Computation Laboratory of Harvard University, and more recently jointly undertaken by Harvard and by the Department of Computer Science of Cornell University.

From the outset, the design of automatic information systems was of principal concern, and the research dealt specifically with the evaluation of a variety of fully automatic methods for information analysis and search. This work resulted in the design of an experimental, fully automatic document retrieval system, called SMART, operating on an IBM 7094 computer, and described in detail in two previous reports in this series, numbered ISR-7 dated June 1964, and ISR-9 dated August 1965.

The SMART system is characterized by the fact that documents and search requests are handled in the natural language without any prior manual analysis, and are processed by one of many different content analysis procedures incorporated into the system. Among these are various statistical and syntactic language analysis methods, and table look-up routines based on a variety of dictionaries and thesauruses. The dictionaries are normally constructed not by committees of subject experts, but semi-automatically starting with representative document collections for each subject area. Since it is unreasonable to expect that the documents retrieved by a single search of the collection should provide adequate answers to all users in all circumstances, iterative search procedures have been used in conjunction

with the SMART system which make it possible to obtain improvements in subsequent searches, using feedback information supplied by the users as a result of earlier searches.

Evaluation results comparing the effectiveness of some of the automatic analysis and search procedures incorporated into the SMART system were first published in report ISR-8 in this series, dated December 1964. More extensive evaluation output is included in the present report, summarizing the work performed during the fall of 1965 and the first half of 1966.

The present report contains work in three main subject areas : automatic and semi-automatic dictionary construction, evaluation output based on results obtained by processing four document collections in three subject areas, and iterative search experiments based on user feedback.

Section I by G. Salton contains a short report on the present state of the SMART project, including also a summary of the research proposed for the immediate future. A complete set of operating instructions for the present version of the SMART system is presented in section II by M. Lesk. A study of this section should make it possible to other interested parties to run portions of the SMART system on different 7094 installations.

Various aspects of the automatic dictionary construction problem are described in sections III, IV and VIII of the present report. Section III by G. Hochgesang contains a description of a very fast concordance generating program which produces keyword-in-context (KWIC) type output from ordinary text input. This program is used to generate the concordances which are later incorporated in the dictionary construction system.

The concordance program described in section III is presently being distributed through the SHARE organization.

In section IV by M. Lesk and G. Salton the complete information dissemination process is examined with emphasis on the use and construction by automatic or semi-automatic techniques of synonym dictionaries and hierarchical subject arrangements. One specific proposal for the fully automatic construction of subject hierarchies is presented in section VIII by G. Blomgren, A. Goodman, and L. Kelly. It is shown in particular how the structure of the hierarchical arrangement changes as various parameters are changed.

Section V of this report by M. Lesk and G. Salton contains in summary form the systems evaluation output produced by the SMART system, based on extensive operations with four document collections in three subject fields (documentation, computer science, and aerodynamics). One document collection used in the experiments consists of document abstracts manually indexed by trained indexers, thus permitting a comparison between the effectiveness of the standard keyword matching techniques and the automatic analysis procedures incorporated into the SMART system. Another collection was available in the form of abstracts as well as longer summaries, thus permitting an evaluation of the effects of document length.

Three sections are devoted to a study of iterative search techniques and user feedback techniques, including sections VI, VII, and IX. Section VI by W. Riddle, T. Horwitz, and R. Dietz examines the effectiveness of a variety of relevance feedback procedures in which the users supply to the system relevance judgments about documents previously retrieved. These judgments are then used automatically to generate a new search request more indicative of user need and preference. This relevance feedback process, first described in detail in section III of report ISR-10, is shown to be extremely effective, and to provide continued improvements in search effectiveness through at least three feedback iterations.

Section VII by V. R. Lesser, and section IX by J. D. Broffitt, H. L. Morgan, and J. V. Soden describe variations of the multi-level search techniques originally introduced in section IV of report ISR-10. These techniques drastically reduce the number of comparisons needed between incoming search requests and stored documents by grouping the stored items, and comparing search requests at first only with a typical item for each group. Individual comparisons are then made only for those documents included in highly scoring groups. The effectiveness of a variety of document grouping procedures used as part of a multi-level search process is evaluated in section IX. Section VII, on the other hand, describes an experiment in which requests previously processed by the system are grouped, rather than documents, and new incoming requests are first compared with these request clusters. The search procedure for a new request is then made to depend on the results obtained with similar requests processed by the system at some previous time. The results of section VII indicate that this heuristic process is useful in reducing search time when the requests to be processed fall into definite patterns, as they may be expected to do in an operational situation.

The last section, number X by M. Lesk contains system design specifications for a SMART type system operating in a time-sharing environment where many users have access to a central document file, and users originate search requests asynchronously, and independently of each other. Equipment specifications and timing considerations are included.