

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



Copyright, 1966
by Cornell University

Use, reproduction, or publication, in whole or in part, is permitted for
any purpose of the United States Government.

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

	Page
SUMMARY	xiii

SECTION I

SALTON, G.: "The SMART System -- Retrieval Results and Future Plans"

1. Introduction	I-1
2. Experimental Results	I-3
3. Discussion and Future Plans	I-5

SECTION II

LESK, M. E.: "Operating Instructions for the SMART Text Processing and Document Retrieval System"

1. Introduction	II-1
1.1. Processing Summary	II-2
1.2. Operating Programs	II-4
2. Basic Operating Procedures	II-4
2.1. Run Outline	II-4
2.2. Tape Setup	II-6
2.3. Input Deck Setup	II-7
3. Specifications for the SMART Retrieval System	II-7
3.1. Specifications Affecting Lookup	II-9
3.2. Specifications Affecting Phrase Searching	II-10

TABLE OF CONTENTS (continued)

	Page
SECTION II (continued)	
3.3. Vector Expansions by Means of Concept- Concept Correlation	II-12
3.4. Vector Expansion by Means of Concept Hierarchies	II-15
3.5. Vector Formation	II-16
3.6. Request-Document Correlation	II-17
3.7. Document-Document Expansion	II-19
3.8. Other Specifications	II-19
4. Data Input	II-20
4.1. Natural Language Documents	II-20
4.2. Binary Documents	II-28
4.3. DOCTAPS	II-29
4.4. Relevance Judgment Data	II-29
4.5. Other Instruction Cards	II-30
5. Tape Preparation Programs	II-31
5.1. Writing a New Library Tape	II-31
5.1.1. Thesaurus and Suffix List Formation	II-33
5.1.2. Statistical Phrase Dictionary	II-35
5.1.3. Syntactic Suffix List	II-37
5.1.4. The Condensed Grammar File	II-38
5.1.5. The Criterion Tree File	II-40
5.1.5.1. Criterion Tree Input Format	II-41
5.1.6. Hierarchy	II-50
5.2. The Document Tape	II-51
5.3. The Program Tape	II-52

TABLE OF CONTENTS (continued)

	Page
SECTION II (continued)	
6. Auxiliary Programs for Use with the SMART System	II-52
6.1. THES	II-54
6.2. MORVAL	II-54
6.3. SOCCER	II-55
7. A Sample Input Deck	II-56
8. Miscellaneous	II-59
8.1. Size Limits	II-59
8.2. Timing	II-59
9. Acknowledgments	II-60

SECTION III

HOCHGESANG, G. T.: "SOCCER - A Concordance Program"

1. Introduction	III-1
2. The Concordance	III-2
A. Definitions	III-2
B. The Input Text	III-2
C. Processing the Text	III-2
D. The Output Format	III-6
3. Tape Usage	III-6
A. Control Cards	III-6
B. The INPUT, OUTPUT, and SMRTAP Tapes	III-6

TABLE OF CONTENTS (continued)

Page

SECTION III (continued)

C.	Scratch Tapes	III-7
4.	Control Cards	III-8
5.	Examples of SØCCER Usage	III-11
6.	Subroutines used by SØCCER	III-13
A.	INØT	III-13
B.	SPECTR	III-15
C.	CLØCK	III-17
7.	Some Details about the SØCCER Program	III-17
A.	Source Deck Changes	III-17
B.	Timing	III-18
Appendix	III-19

SECTION IV

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

1.	Introduction	IV-1
2.	Language Analysis	IV-2
3.	Dictionary Construction	IV-7
A)	The Synonym Dictionary (Thesaurus)	IV-7
B)	The Null Thesaurus and Suffix List	IV-15
C)	The Phrase Dictionaries	IV-21

TABLE OF CONTENTS (continued)

Page

SECTION IV (continued)

D)	The Concept Hierarchy	IV-27
4.	Dictionary Performance	IV-32
A)	The Null Thesaurus	IV-33
B)	The Regular Thesaurus	IV-38
C)	The Phrase Dictionary	IV-42
5.	Automatic Thesaurus Construction	IV-44
A)	Fully-Automatic Methods	IV-48
B)	Semi-Automatic Methods	IV-50
C)	Sample Thesaurus Generation	IV-56
6.	Semi-Automatic Hierarchy Formation	IV-59

SECTION V

LESK, M. E., and SALTON, G.: "Design Criteria for Automatic Information Systems"

1.	Introduction	V-1
2.	The SMART Experiments	V-3
3.	Evaluation Results and Design Criteria	V-11
A)	Indexing Depth and Document Length	V-11
B)	Synonym Recognition	V-16
C)	Phrase Processing	V-19
D)	Statistical Association Methods	V-22

TABLE OF CONTENTS (continued)

Page

SECTION V (continued)

E)	Hierarchical Subject Expansion	V-28
F)	Manual Indexing	V-30
G)	Iterative Searching	V-32
H)	Summary	V-33

SECTION VI

RIDDLE, W., HORWITZ, T., and DIETZ, R.: "Relevance Feedback in an
Information Retrieval System"

1.	Introduction	VI-1
2.	Principal Methods	VI-4
A)	Determination of the Number of Documents Retrieved	VI-5
B)	The Effect of the Correlation Function	VI-5
C)	Determination of the Relevance Weighting Factors	VI-6
D)	Determination of the Value of α	VI-8
E)	Termination of the Modification Process	VI-10
3.	Experimental Results	VI-10
4.	Conclusions	VI-12
Appendix A		VI-16
Appendix B (by. E. M. Keen)		VI-19

TABLE OF CONTENTS (continued)

Page

SECTION VII

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

1. Introduction	VII-1
2. A Modified Clustering Algorithm and a Corresponding Two-Level Search Strategy	VII-3
3. Advantages of the Query Clustering System	VII-5
4. Design of an Experiment to Compare the Modified with the Normal Two-Level Search Scheme	VII-7
A) Problem Areas	VII-7
B) Tests to Compare the Effectiveness of Each Search Procedure	VII-7
C) Implementation of the Normal and Modified Two-Level Search Schemes	VII-9
D) Test Data Base	VII-12
5. Actual Comparisons of the Modified versus the Normal Two-Level Searches	VII-13
A) Data Generated for Two-Level Search Algorithm	VII-13
B) Data Generated for Modified Two-Level Search Algorithm	VII-14
C) Experimental Evaluation	VII-16
D) Evaluation Results	VII-25
6. A New Criterion for Search Effectiveness	VII-27
7. Conclusions	VII-28
Appendix A	VII-30

TABLE OF CONTENTS (continued)

Page

SECTION VIII

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

1. Introduction	VIII-1
2. Automatic Construction of Hierarchies	VIII-2
3. Outline of the Investigation	VIII-12
Appendix A	VIII-15

SECTION IX

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

1. Introduction	IX-1
2. Similarity Measures	IX-4
3. Rocchio's Procedure	IX-5
4. Bonner's Procedure	IX-7
5. The Experiment	IX-10
6. Evaluation	IX-11
7. Results and Conclusions	IX-12

TABLE OF CONTENTS (continued)

Page

SECTION X

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

1. Introduction	X-1
2. Principles	X-2
3. Methods	X-5
4. Practicalities	X-9
5. Conclusions	X-17

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.

