

Department of Computer Science

Cornell University

Ithaca, New York 14850

Scientific Report No. ISR-14

INFORMATION STORAGE AND RETRIEVAL

to

The National Science Foundation

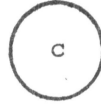
Reports on Analysis, Search and Iterative Retrieval

Ithaca, New York

October 1968

Gerard Salton

Project Director



Copyright 1968
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

Kenneth M. Brown
Robert L. Constable
Richard W. Conway
Barbara Evers
Sally Grove
Juris Hartmanis
John E. Hopcroft
Howard L. Morgan
Joann Newman
Rosalind Pasquali
Christopher Pottle
Gerard Salton
Alan C. Shaw
Stephen Stephenson
Roland A. Sweet
Robert A. Wagner
Robert J. Walker
Peter Wegner
Donna Williamson
Robert E. Williamson
William S. Worley

Project Staff in the Division of Engineering and Applied Physics

Harvard University

Jeffrey Bean
Jeffrey Golden
Michael Lesk
E. Ricardo Quinones

REPORTS ON ANALYSIS, SEARCH AND ITERATIVE RETRIEVAL

TABLE OF CONTENTS

	Page
SUMMARY	xiii

PART ONE

SYSTEM DESIGN

I. LESK, M. E.

"Design of a Revised On-Line Information Retrieval System"

1. Introduction	I-1
2. Supervisor Organization	I-4
3. System Procedures	I-11
A) Request and Text Input	I-13
B) Request and Text Lookup	I-14
C) Automatic Thesaurus Processing	I-27
D) Phrase Processing	I-30
E) Hierarchical Processing	I-34
F) Concept Vector Formation and Storage	I-36
G) Searching of Document Collections	I-39
H) Clustering of Document Collections	I-43
I) Relevance Feedback	I-43
J) Dictionary Displays	I-44
K) Citation Searching	I-47
L) Class Information	I-49
M) Selective Information Dissemination	I-49
N) User Information Files	I-50
4. Equipment	I-50
5. Summary	I-52

TABLE OF CONTENTS (continued)

	Page
I. continued	
References	I-54
Appendix	I-56
II. WILLIAMSON, D.	
"The Cornell Implementation of the SMART System"	
Abstract	II-1
1. Introduction	II-1
2. Basic Cornell System Organization	II-2
3. The SMART System Routines	II-12
A) Control Routines	II-12
B) Inner System Routines	II-15
References	II-20

PART TWO

ANALYSIS AND SEARCH

III. LESK, M. E. AND SALTON, G.	
"Relevance Assessments and Retrieval System Evaluation"	
Abstract	III-1
1. Introduction	III-1
2. The Relevance Problem	III-3
3. The Experiment	III-8
4. Experimental Results	III-14
5. Judgment Consistency and Performance Measures	III-23
6. Machine Search Effectiveness	III-28

TABLE OF CONTENTS (continued)

	Page
III. continued	
References	III-34
Appendix	III-36
IV. COYAUD, M.	
"Resolution of Lexical Ambiguities in Ophthalmology"	
1. Introduction	IV-1
2. Procedures for Devising the Polysemy Rules .	IV-2
A) The Rules Inspired by Corpus A . .	IV-3
B) Notes to the Polysemy Rules . . .	IV-8
C) The Control of the Rules by Corpus B .	IV-11
3. Conclusion	IV-13
References	IV-15
Annex I	IV-16
V. DATTOLO, R. T.	
"A Fast Algorithm for Automatic Classification"	
Abstract	V-1
1. Introduction	V-1
2. The N^2 Problem	V-2
3. Doyle's Algorithm	V-3
4. Satisfaction of Termination Condition . .	V-7
A) Non-convergence of Doyle's Algorithm .	V-7
B) Termination of Modified Algorithm . .	V-10
5. Implementation	V-13

TABLE OF CONTENTS (continued)

	Page
V. continued	
6. Experimental Results	V-16
A) The Scoring Function	V-16
B) Movement of Documents	V-18
C) Initial Clusters	V-24
D) Evaluation of Results	V-24
7. Conclusion	V-28
References	V-31
VI. SALTON, G. AND WILLIAMSON, D. K.	
"A Comparison Between Manual and Automatic Indexing Methods"	
Abstract	VI-1
1. Introduction	VI-1
2. The Evaluation of Information Systems	VI-3
3. The Test Design	VI-8
A) The MEDLARS Evaluation Study	VI-8
B) Design of the SMART Test	VI-11
4. SMART-MEDLARS Comparison	VI-19
5. Comparison of SMART Analysis Methods	VI-22
6. Conclusions	VI-28
References	VI-34
Appendix A	VI-36
Appendix B	VI-43

TABLE OF CONTENTS (continued)

Page

PART THREE

USER FEEDBACK PROCEDURES

VII. SALTON, G.

"Search and Retrieval Experiments in Real-Time Information Retrieval"

Abstract	VII-1
1. Introduction	VII-1
2. Performance Characteristics of Information Systems	VII-3
3. User Feedback Retrieval Methods	VII-12
A) General Methodology	VII-12
B) Positive Feedback	VII-14
C) Negative Feedback	VII-19
References	VII-29

VIII. IDE, E.

"New Experiments in Relevance Feedback"

Abstract	VIII-1
1. The Relevance Feedback Procedure	VIII-1
2. The Experimental Environment	VIII-4
3. Earlier Results in the Same Environment	VIII-6
4. Evaluation of Retrieval Performance	VIII-7
A) The "Feedback Effect" in Evaluation	VIII-7
B) Performance Measures	VIII-10
C) Statistical Tests	VIII-12

TABLE OF CONTENTS (continued)

	Page
5. Experimental Results	VIII-14
A) Two Strategies Using Relevant Documents Only	VIII-15
B) Varying the Amount of Feedback	VIII-16
C) Strategies Using Nonrelevant Documents	VIII-20
6. Summary and Recommendations	VIII-28
References	VIII-30

IX. LESK, M. E. AND SALTON, G.

"Interactive Search and Retrieval Methods Using Automatic Information Displays"

Abstract	IX-1
1. Introduction	IX-1
2. Fully-Automatic Retrieval	IX-4
3. User Interaction Through Pre-Search Methods	IX-8
4. User Interaction Through Post-Search Methods	IX-13
5. Evaluation Results and Discussion	IX-16
A) Recall-Precision Results	IX-17
B) Overall Evaluation	IX-20
6. Conclusion	IX-31
References	IX-35

X. DAVIS, M. C., LINSKY, M. D., AND ZELKOWITZ, M. V.

"A Relevance Feedback System Employing a Dynamically Evolving Document Space"

Abstract	X-1
1. Introduction	X-1

TABLE OF CONTENTS (continued)

	Page
X. continued	
2. Proposed Study	X-5
3. Experimental Results	X-9
4. Results and Conclusions	X-23
References	X-25
Appendix	X-26
 XI. BRAUEN, T. L., HOLT, R. C., AND WILCOX, T. R.	
"Document Indexing Based on Relevance Feedback"	
Abstract	XI-1
1. Introduction	XI-1
2. Method	XI-3
3. The Experiment	XI-6
4. Experimental Results	XI-8
5. Discussion	XI-12
References	XI-15
Appendix	XI-16
 XII. BORODIN, A., KERR, L., AND LEWIS, F.	
"Query Splitting in Relevance Feedback Systems"	
Abstract	XII-1
1. Introduction	XII-1
2. The Query Splitting Algorithm	XII-3
3. Evaluation and Results	XII-5
4. Conclusions and Suggestions for Further Research	XII-14

TABLE OF CONTENTS (continued)

	Page
References	XII-15
Appendix	XII-16
1. Introduction	XII-16
2. General Algorithm	XII-16
3. System Operation	XII-17

XIII. CRAWFORD, R. G., AND MELZER, H. Z.

"The Use of Relevant Documents Instead of Queries in Relevance Feedback"

Abstract	XIII-1
1. Introduction	XIII-1
2. Motivation and Assumptions	XIII-3
3. Implementation	XIII-6
4. Results	XIII-8
A) Feedback Using Only Relevant Documents	XIII-8
B) Source Document Used as Original Query	XIII-18
5. Conclusions	XIII-21
References	XIII-26

PART FOUR

EDITING PROGRAMS

XIV. BEAN, JEFFREY

"Bean's Automatic Tape Manipulator - A Description, and Operating Instructions"

1. General Description	XIV-1
2. Operating Instructions	XIV-2

TABLE OF CONTENTS (continued)

	Page
XV. QUINONES, RICARDO, E.	
"EDIT - An Editing Subroutine"	
1. General Description	XV-1
2. Requirements and Specifications for BATMAN .	XV-2
3. EDIT Control Card Format	XV-5
A) Serialization	XV-5
B) Editing Commands	XV-7
C) Special Commands	XV-13
4. Editing Principles	XV-20
A) Temporary Cards ("'" and D Cards) . .	XV-20
B) Permanent Cards (P and X Cards). . .	XV-21
C) String Contents	XV-22
5. Miscellany and Tables	XV-22