

TABLE OF CONTENTS

SUMMARY	Page vii
-------------------	-------------

PART ONE

SMART SYSTEM DESIGN

I. WILLIAMSON, D., WILLIAMSON, R., AND LESK, M.

"The Cornell Implementation of the SMART System"

Abstract	I-1
1. Introduction	I-1
2. Basic System Organization	I-2
A) Input of Printed Text	I-3
B) Document Clustering for Search Purposes	I-5
C) The Selection of Documents to be Searched	I-14
D) The Searching of the Document Groups	I-28
E) Search Evaluation	I-41
3. Access to the SMART System	I-54
4. Basic SMART System Flowchart	I-56
References	I-62

II. MURRAY, D.

"A Scatter Storage Scheme For Dictionary Lookups"

1. Introduction	II-1
2. Basic Scatter Storage	II-2
A) Method	II-2
B) Collisions	II-3
C) Table Layout and Search Procedure	II-4

SMART Project Staff

Robert Crawford

Barbara Evers

Marcia Kerchner

Michael Lesk (Harvard)

Harry Melzer

Rosalind Pasquali

Jacob Razon

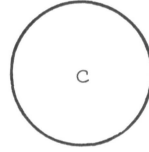
Angie Rettig

Gerard Salton

Donna Williamson

Robert Williamson

Steven Worona



Copyright 1969
by Cornell University

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

Department of Computer Science

Cornell University

Ithaca, New York 14850

Scientific Report No. ISR-16

INFORMATION STORAGE AND RETRIEVAL

to

The National Science Foundation

Ithaca, New York

September 1969

Gerard Salton

Project Director

TABLE OF CONTENTS (continued)

	Page
II. continued	
D) Theoretical Expectations	II-5
3. Virtual Scatter Storage	II-10
A) Method	II-10
B) Collision Problem	II-12
4. Experiments with Algorithms for Generating Hash Addresses	II-15
A) Dictionaries	II-15
B) Hash Coding Algorithms	II-17
C) Evaluation	II-21
5. A Practical Lookup Scheme	II-32
A) General Description	II-32
B) Table Layout	II-33
C) Search Considerations	II-34
D) Performance	II-38
E) Comparisons	II-40
6. Extensions	II-40
A) Larger Dictionaries	II-40
B) Suffix Removal	II-41
7. Conclusions	II-41
References	II-42

III. JOINER, J. AND WERNER, L.

"A New Evaluation Measure"

Abstract	III-1
1. Introduction	III-1
2. Problems of Evaluation	III-2
3. Criteria for a Good Evaluation Measure	III-4

TABLE OF CONTENTS (continued)

Page

III. continued

4. The Probability Measure	III-5
5. Tests	III-8
Bibliography	III-11

PART TWO

CONTENT ANALYSIS METHODS

IV. SALTON, G.

"Automatic Processing of Foreign Language Documents"

Abstract	IV-1
1. Introduction	IV-1
2. The SMART System	IV-3
3. The Evaluation of Language Analysis Methods	IV-6
4. Multi-lingual Thesaurus	IV-10
5. Foreign Language Retrieval Experiment	IV-12
6. Failure Analysis	IV-19
7. Conclusion	IV-25
References	IV-28
Appendix	IV-29

V. WEISS, S. F.

"Syntax in Text Analysis"

Abstract	V-1
1. Introduction	V-1

TABLE OF CONTENTS (continued)

	Page
V. continued	
2. Statistical Phrases	V-2
3. Syntactic Phrases	V-4
4. Cooccurrence	V-5
5. Elimination of the Phrase List	V-9
6. Analysis of Results	V-11
7. Conclusion	V-16
References	V-18
VI. WEISS, S. F.	
"Template Analysis and its Application to Natural Language Processing"	
Abstract	VI-1
1. The Basics of Template Analysis	VI-1
A) Introduction	VI-1
B) Types of templates	VI-4
C) Applicability of template analysis	VI-7
2. An Implementation of Natural Language Analysis by Template Analysis	VI-8
A) Keyword Analysis	VI-9
B) Implementation conventions	VI-13
3. An Implementation of Template Analysis	VI-26
A) Date phrases	VI-27
B) Journal phrases	VI-32
C) Author phrases	VI-35
D) Experiments and results	VI-41
E) Conclusion	VI-44
Bibliography	VI-46

TABLE OF CONTENTS (continued)

	Page
VI. continued	
Appendix A	VI-47
Appendix B	VI-50
VII. FAITH, B. AND JENSEN, J.	
"The Combination of Thesaurus and Word Form Vectors"	
Abstract	VII-1
1. Introduction	VII-1
2. Procedure	VII-2
3. Results	VII-3
4. Further Studies	VII-10
References	VII-11
VIII. McNEIL, J. W., AND WETHERELL, C. S.	
"Bibliographic Data as an Aid to Document Retrieval"	
Abstract	VIII-1
1. Introduction	VIII-1
2. The Experiment	VIII-3
3. The Statistical Measure	VIII-8
4. The Results	VIII-11
5. Conclusions	VIII-12
References	VIII-15

TABLE OF CONTENTS (continued)

Page

PART THREE

USER FEEDBACK PROCEDURES

IX. BROWN, J. S., AND REILLY, P. D.

"The Use of Statistical Significance in Relevance Feedback"

Abstract	IX-1
1. Introduction	IX-1
2. Query Construction	IX-10
3. Conduct of the Experiment	IX-13
4. Experimental Results	IX-14
5. Conclusions and Recommendations	IX-33
References	IX-35
Appendix A	IX-37

X. CIRILLO, C., CHANG, Y. K., AND RAZON, J.

"Evaluation of Feedback Retrieval using Modified Freezing, Residual Collection and Test and Control Groups"

Abstract	X-1
1. Introduction	X-1
Part A: Evaluation of Feedback Retrieval Using Modified Freezing	X-3
1. Introduction	X-4
2. Modified Freezing	X-4
3. Evaluation Results	X-7
4. Discussion	X-8

TABLE OF CONTENTS (continued)

	Page
X. continued	
Part B: Evaluation of Feedback Retrieval Using Residual Collection Feedback	X-11
1. Statement of the Problem	X-12
2. Summary of Methods	X-12
3. Results and Conclusions	X-14
Part C: Evaluation of Feedback Retrieval Using Test and Control Groups	X-22
1. Introduction	X-23
2. Process Description	X-23
3. Experimental Results and Evaluation	X-27
4. Conclusions	X-33
References	X-34
XI. IDE, E., AND SALTON, G.	
"Interactive Search Strategies and Dynamic File Organization in Information Retrieval"	
Abstract	XI-1
1. Retrieval System Performance	XI-1
2. Request Space Modifications	XI-4
A) Relevance Feedback	XI-4
B) Positive and Negative Strategies	XI-6
C) Selective Negative Feedback	XI-19
3. Document Clustering	XI-22
4. Document Space Modification	XI-28
5. Conclusion	XI-32
References	XI-33

TABLE OF CONTENTS (continued)

Page

XII. LEVENTHAL, T., AND MILLER, R.

"Query Splitting Using Relevant Documents Instead of Queries in Relevance Feedback"

Abstract	XII-1
1. Introduction	XII-1
2. Motivations and Assumptions	XII-3
3. Implementation	XII-5
4. Evaluation and Results	XII-8
5. Conclusions	XII-13
References	XII-14

PART FOUR

CLUSTERING METHODS

XIII. DATTOLA, R.

"Experiments with a Fast Algorithm for Automatic Classification"

Abstract	XIII-1
1. Introduction	XIII-1
2. General Description	XIII-2
3. Implementation	XIII-6
A) Initial Clusters	XIII-7
B) Overlap	XIII-8
C) Algorithm	XIII-11
4. Evaluation	XIII-15
A) Evaluation Measures	XIII-16
B) Internal Evaluation	XIII-24

TABLE OF CONTENTS (continued)

Page

XIII. continued

C) Initial Clusters	XIII-23
D) Number of Clusters	XIII-36
E) Overlap	XIII-42
F) Cutoff	XIII-45
G) Percent Loose Clustered	XIII-48
H) External Evaluation	XIII-51
5. Conclusion	XIII-59
References	XIII-62

XIV. RIEBER, S., AND MARATHE, V. P.

"The Single Pass Clustering Method"

Abstract	XIV-1
1. Introduction	XIV-1
2. The Program	XIV-4
3. Investigation and Results	XIV-7
A) Correlation Comparison	XIV-7
B) Disjoint-Overlapping Comparison	XIV-8
C) Variation of Document Order	XIV-10
D) SMART Evaluation	XIV-10
4. Conclusions	XIV-14
References	XIV-18
Appendix 1	XIV-19
Appendix 2	XIV-27

TABLE OF CONTENTS (continued)

Page

XV. WORONA, S.

"Query Clustering in a Large Document Space"

Abstract	XV-1
1. Introduction	XV-1
2. Generating Clusters	XV-2
3. Searching Clustered Collections	XV-5
4. Parameters for Evaluating Cluster Searches	XV-5
5. The Experiment	XV-8
6. Results	XV-12
References	XV-15
Appendix A	XV-17
Appendix B	XV-22