

08

**THE COMPUTATION LABORATORY
OF
HARVARD UNIVERSITY**



INFORMATION STORAGE AND RETRIEVAL

**Scientific Report No. ISR-9
to
The National Science Foundation**

**Cambridge, Massachusetts
August 1965**

**Gerard Salton
Project Director**

T H E C O M P U T A T I O N L A B O R A T O R Y

Harvard University
Cambridge, Massachusetts

Scientific Report No. ISR-9
INFORMATION STORAGE AND RETRIEVAL
to
The National Science Foundation

Cambridge, Massachusetts

August 1965

Gerard Salton

Project Director



Copyright, 1965

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

STAFF OF THE COMPUTATION LABORATORY AND OF THE COMPUTING CENTER

Addison, Lynda	Kuno, Susumu
Arkin, Lois	Lemieux, Marguerite
Aymond, George	Lemmon, Alan
Bartee, Thomas	Lesk, Michael
Benincasa, Edward	Lewis, Thomas
Bossert, William	Lewko, Joseph
Bossin, Benjamin	Lynch, Mary
Burkhart, Donald	Lyons, Edward
Burns, Robert	Mahony, Rita
Cane, Mark	McAnulty, Michael
Canter, Jane	McCarter, Calvin
Carpenter, George	McDonald, John
Chuntaketta, Somwong	McFarland, Forrest
Coffin, Howard	McLean, Margaret
Collins, Irene	McMahon, Constance
Cone, Evelyn	McManus, George
Conway, Paul	Murphy, Thomas J.
Corbato, Isabel	Nathanson, Sheila
D'Agostino, Jean	Newell, John
Dauway, Shervington	O'Brien, Francis
Delery, Richard	Oettinger, Anthony
Dingler, Richard	Parry, Edward, Jr.
Dolan, Arthur	Perkins, Donald
Donahue, James	Powers, Richard
Drew, David	Priver, Arthur
Droge, H. Alfred	Prowse, James
Eckian, Judith	Puckett, Allan
Evslin, Tom	Razar, Michael
Gaetano, Leonard	Robbins, Mary Jane
Geary, Robert	Rocchio, Joseph
Greibach, Sheila	Rudy, Jeanette
Haitovsky, Yoel	Ruyle, Adrian
Hassell, Marie	Salton, Gerard
Harris, Claudine	Sanborn Sill, Jacquelin
Hazel, Inez	Shapiro, George
Hochgesang, Guy	Smith, Carol
Hubbard, John	Stockson, Eda
Johnson, Shirley	Taylor, Ethel
Kirsch, Elena	Thorpe, Rodney

STAFF OF THE COMPUTATION LABORATORY AND OF THE COMPUTING CENTER (continued)

Thompson, Howard
Tingey, Thomas
Tukis, Cynthia
Van Bemmell, John
Wang, Hao

Whalen, Richard
Winiecki, Kenneth
Wooton, Thomas
Yarbrough, Lynn
Young, Stephen
Zachary, Norman

TABLE OF CONTENTS

	Page
SUMMARY	xvi

SECTION I

SALTON, GERARD: "A Progress Report on SMART"

1. Introduction	I-1
2. The SMART Project (1964 - 1965)	I-3
3. Future Projects	I-6

SECTION II

EVSLIN, TOM: "A General Discussion"

1. Introduction	II-1
2. The LOOKUP and Phrase Detection	II-1
3. Further Modifications to the Document Vectors	II-6
A. Concept-Concept Correlation	II-6
B. The Hierarchy	II-8
4. Correlation	II-9
5. Document-Document Correlation to Expand the Answer List	II-11
6. Evaluation	II-12
7. Graphical System Summary	II-13

TABLE OF CONTENTS (continued)

Page

SECTION III

EVSLIN, TOM: "A Specific Description of the New SMART System"

1.	Operating Instructions	III-1
	A. Input Parameters	III-1
	B. Control Cards	III-5
	C. English Text Following *FIND, *LIKE, and *TEXT Cards	III-7
	D. Evaluation Prediction Decks	III-7
	E. Order of Input Decks	III-8
2.	System Description	III-9
3.	System Flowchart	III-13
4.	TAPE Assignments	III-18

SECTION IV

LESK, MICHAEL: "Specification Input to the Revised SMART System"

1.	Introduction	IV-1
2.	Parameter Names Recognized by the SMART System	IV-2
	A. Parameters which Affect the Flow of Control.	IV-2
	B. Parameters which Set Weights, Cutoffs, etc.	IV-3

TABLE OF CONTENTS (continued)

Page

SECTION IX (continued)

4. Termination of Processing by UPCRIT	IX-4
5. An Example of Criterion Tree Updating	IX-5
6. Miscellaneous Data about UPCRIT	IX-6
Appendix A	IX-7

SECTION X

LEMMON, ALAN: "The Revised Syntactic Processing for SMART"

1. Introduction	X-1
2. Use and Function of System Routines	X-1
A. BTOKUN (Binary-TO-KUN Converter)	X-1
B. CRITER (CRITERion tree routine)	X-3
C. TRECND (TREe maker, CoNDensed format)	X-6
3. Technical Considerations	X-12
4. Summary	X-19
Appendix 1	X-20
Appendix 2	X-22

TABLE OF CONTENTS (continued)

Page

SECTION XI

PROUSE, JAMES: "Syntactic Analysis of Incomplete Sentences in the
SMART System"

1.	Introduction	XI-1
2.	A Brief Description of the Analyzer	XI-2
3.	Methods for Solving the Acceptance Problem	XI-3
4.	Analyzer Modifications	XI-6

SECTION XII

RAZAR, MICHAEL: "The Tree Matching Program-MATCH"

1.	Introduction	XII-1
2.	General Considerations	XII-3
3.	Calling Sequence for MATCH; Returning from MATCH	XII-10

SECTION XIII

EVSLIN, TOM, HOCHGESANG, GUY and LESK, MICHAEL: "Vector Merging and Vector
Correlations"

1.	Introduction	XIII-1
2.	The SUPER 8 Routine	XIII-2
	A. Description	XIII-3
	B. SUPER 8 Flowchart	XIII-5

TABLE OF CONTENTS (continued)

	Page
SECTION XIII (continued)	
3. The MERVEC Routine	XIII-10
A. Description	XIII-10
B. MERVEC Flowchart	XIII-13
C. Internal Subroutines	XIII-16
4. The WRITE Routine	XIII-19
A. Description	XIII-19
B. WRITE Flowchart	XIII-21
5. The SUPPLY and SUPTAP Subroutine	XIII-22
A. Description	XIII-22
B. SUPPLY Flowchart	XIII-24
6. The ERROR Routine	XIII-25

SECTION XIV

CANE, MARK and SHAPIRO, GEORGE: "Concept-Concept and Document-Document
Correlations"

1. Introduction	XIV-1
2. Correlation Algorithms	XIV-2
3. Concept-Concept Correlations	XIV-12
4. Document-Document Correlations	XIV-13

TABLE OF CONTENTS (continued)

Page

SECTION XV

RAZAR, MICHAEL and SHAPIRO, GEORGE: "Hierarchy Set-Up and Hierarchy and
Concept-Concept Expansion Procedures"

1.	Introduction	XV-1
2.	Card and Tape Formats; Size Limitations	XV-2
3.	Description of Algorithms Used by SETUP	XV-7
4.	The Expansion Programs	XV-11

SECTION XVI

HOCHCESANG, GUY: "The Reports Format"

1.	Introduction	XVI-1
2.	Formats for Reports	XVI-1

SECTION XVII

LESK, MICHAEL: "Evaluation of Retrieval Results in the Extended
SMART System"

1.	Introduction	XVII-1
2.	Single-run Evaluation	XVII-6
3.	Multiple-run Evaluation	XVII-10

TABLE OF CONTENTS (continued)

Page

SECTION XVIII

LESK, MICHAEL: "A Program to Evaluate the Iterative Search
Process (MORVAL)"

SECTION XIX

ROCCHIO, JOSEPH: "Combinations of Analysis Methods - The Merged
Output Results"

1.	Introduction	XIX-1
2.	The Merging Process	XIX-2
3.	Sample Results	XIX-4
4.	Conclusions	XIX-10

SECTION XX

HARRIS, CLAUDINE: "Analysis of Student Requests"

	Introduction	XX-1
1.	Description of the Student Requests	XX-2
2.	Retrieval Results	XX-3
3.	Combined Processing Methods	XX-11
4.	Revision of Requests and Dictionaries	XX-15

TABLE OF CONTENTS (continued)

Page

SECTION XXI

ROCCHIO, JOSEPH: "Evaluation Viewpoints in Document Retrieval"

Summary	XXI-1
1. Introduction	XXI-1
2. Macro Evaluation	XXI-3
3. Micro Evaluation	XXI-4
4. Example	XXI-5
5. Conclusion	XXI-10

SECTION XXII

SILLERS, S. J.: "Distinguishing Retrieved From Nonretrieved Information:
The Cut-off Problem"

1. The Cut-off Problem	XXII-1
2. Possible Methods of Solution	XXII-1
3. The Computer Program	XXII-4
4. Data Sets Examined	XXII-6
5. Results and Conclusions	XXII-7

TABLE OF CONTENTS (continued)

Page

SECTION XXIII

ROCCHIO, JOSEPH: "Relevance Feedback in Information Retrieval"

Summary	XXIII-1
1. System Model	XXIII-2
2. Request Formulation	XXIII-2
3. Request Optimization	XXIII-4
4. Relevance Feedback	XXIII-7
5. Initial Experimental Results	XXIII-11

SECTION XXIV

ROCCHIO, JOSEPH and SALTON, GERARD: "Information Search Optimization and Iterative Retrieval Techniques"

Abstract	XXIV-1
1. Introduction	XXIV-2
2. The Automatic Dictionary Process	XXIV-3
3. Request Optimization Using Relevance Feedback	XXIV-12
4. Automatic Modification of the Analysis Process	XXIV-19