

Department of Computer Science

Cornell University

Ithaca, New York 14850

Scientific Report No. ISR-12

INFORMATION STORAGE AND RETRIEVAL

to

The National Science Foundation

Reports on Evaluation, Clustering, and Feedback

Ithaca, New York

June 1967

Gerard Salton

Project Director



Copyright, 1967
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
Richard W. Conway
Margaret Dodd
Patrick C. Fischer
Sally Grove
Juris Hartmanis
Eleanor R. Ide
E. Michael Keen
Joann Newman
Gerard Salton
Robert J. Walker
Peter Wegner
Robert E. Williamson

Project Staff in the Division of Engineering and Applied Physics

Harvard University

Jeffrey Bean
Claudine Harris
Sally Hobbs
Michael Lesk
E. Ricardo Quinones

TABLE OF CONTENTS

SUMMARY	Page xi
-------------------	------------

PART ONE

EVALUATION

I. SALTON, G.

"The SMART Project - Status Report and Plans"

1. Introduction	I-1
2. Experimental Results	I-3
3. Future Plans	I-6
A) Additional Evaluation Experiments	I-6
B) New Real-Time Operating System	I-9

II. LESK, M. E.

"SIG - The Significance Programs for Testing the Evaluation Output"

1. Introduction	II-1
2. T-tests	II-3
3. Sign Tests	II-8
4. Program Operation	II-10
5. Output	II-13
References	II-22

TABLE OF CONTENTS (continued)

	Page
III. SALTON, G. and LESK, M. E.	
"Computer Evaluation of Indexing and Text Processing"	
1. Introduction	III-1
2. The SMART System	III-3
A) Basic Organization	III-3
B) Evaluation Process	III-5
C) Significance Computations	III-12
3. Experimental Results	III-22
A) Test Environment	III-22
B) Document Length	III-27
C) Matching Function and Term Weights	III-31
D) Language Normalization - The Suffix Process	III-38
E) Synonym Recognition	III-41
F) Phrase Recognition	III-46
G) Hierarchical Expansion	III-54
H) Manual Indexing	III-55
4. Concluding Comments	III-60
Acknowledgment	III-65
References	III-66

PART TWO

CLUSTER SEARCH METHODS

IV. IDE, E., WILLIAMSON, R. and WILLIAMSON, D.

"The Cornell Programs for Cluster Searching and Relevance Feedback"

TABLE OF CONTENTS (continued)

	Page
IV. continued	
1. Design Criteria	IV-1
2. Basic Cornell System Organization	IV-1
3. System Implementation of Rocchio's Clustering Algorithm	IV-2
4. System Implementation of Relevance Feedback	IV-6
5. Further Details of the Cornell System Organization	IV-8
A) Routines Available Without User Reprogramming	IV-8
B) Service Routines Available to the Experimental Programmer	IV-10
References	IV-13

V. SALTON, G.

"Search Strategy and the Optimization of Retrieval Effectiveness"

Abstract	V-1
1. Introduction	V-1
2. Cluster Search Process	V-4
A) Overall Process	V-5
B) Cluster Generation	V-6
C) Cluster Searching and Evaluation	V-18
3. Relevance Feedback	V-30
A) Overall Process	V-30
B) Feedback Evaluation	V-33

TABLE OF CONTENTS (continued)

	Page
V. continued	
4. Adaptive User-Controlled Multi-Level Search . . .	V-42
References	V-46
VI. GRAUER, R. T. and MESSIER, M.	
"An Evaluation of Rocchio's Clustering Algorithm"	
Abstract	VI-1
1. Introduction	VI-1
2. A Description of Rocchio's Algorithm	VI-2
3. Problems to be Studied	VI-4
4. Evaluation Scheme	VI-6
A) Tabulation of Results	VI-8
B) Detailed Analysis	VI-16
C) Conclusions and Remaining Questions	VI-37
References	VI-39
VII. LEECH, P. C. and MATLACK, R. C. Jr.	
"Information Retrieval: Dictionary Representations and Cluster Evaluation"	
Abstract	VII-1
1. Introduction	VII-1
2. Rocchio's Clustering Procedure	VII-3
3. The Experiment	VII-5

TABLE OF CONTENTS (continued)

	Page
VII. continued	
4. Results and Evaluation	VII-7
A) The Null Dictionary	VII-7
B) The Thesaurus Dictionary	VII-13
C) Comparison of the Null and Thesaurus Dictionaries	VII-14
5. Conclusions	VII-16
References	VII-18

PART THREE

USER FEEDBACK ALGORITHMS

VIII. IDE, E.

"User Interaction with an Automated Information Retrieval System"

Abstract	VIII-1
1. Introduction	VIII-1
2. The Experimental Environment	VIII-5
3. Experimental Results	VIII-7
A) Comparison of the Cranfield and ADI Collections	VIII-8
B) Strategies Using Relevant Documents Only	VIII-10
C) Amount of Feedback	VIII-13
D) Strategies Using Non-relevant Documents	VIII-18
4. Conclusions	VIII-28
Acknowledgments	VIII-31
References	VIII-32

TABLE OF CONTENTS (continued)

	Page
IX. KELLY, J.	
"Negative Response Relevance Feedback"	
1. Introduction	IX-1
2. Principal Algorithm	IX-2
3. Experimental Method	IX-5
4. Results	IX-7
5. Conclusion	IX-9
References	IX-10
Appendix	IX-11
X. FRIEDMAN, S. R., MACEYAK, J. A. and WEISS, S. F.	
"A Relevance Feedback System Based on Document Transformation"	
Abstract	X-1
1. The Problem	X-1
2. The Implementation	X-3
3. Experimental Results	X-7
4. Conclusions	X-17
References	X-19
XI. AMREICH, M., GRISSOM, G., MICHELSON, D., IDE, E.	
"An Experiment in the Use of Bibliographic Data as a Source of Relevance Feedback in Information Retrieval"	
Abstract	XI-1

TABLE OF CONTENTS (continued)

	Page
XI. continued	
1. Introduction	XI-1
2. The Bibliographic Assumptions	XI-2
3. The Problem and Method	XI-3
4. Query Alteration in Feedback	XI-9
5. Evaluation	XI-9
6. Conclusions and Recommendations	XI-25
References	XI-27
XII. HALL, H. A. and WEIDERMAN, N. H.	
"The Evaluation Problem in Relevance Feedback Systems"	
Abstract	XII-1
1. Introduction	XII-1
2. The Present Evaluation System	XII-2
3. The Problem	XII-5
4. The Solution	XII-8
5. Preliminary Results	XII-10
6. Conclusion	XII-16
References	XII-18

Summary

The present report is the twelfth in a series covering research in automatic storage and retrieval conducted by the Department of Computer Science at Cornell University with the assistance of the Division of Engineering and Applied Physics at Harvard University. Unlike some of the preceding reports (ISR 7, 8, 9 and 11), the present report does not deal principally with the programming design of the fully-automatic SMART document retrieval system, now implemented both at Harvard on an IBM 709⁴ computer and at Cornell on a CDC 160⁴. Instead, the report deals with the search and retrieval experiments undertaken over the last year (June 1966 to June 1967), and with the evaluation results obtained by applying the SMART System to document collections in computer science, documentation, and aerodynamics.

The report is divided into three parts, titled Evaluation, Cluster Searching, and User Feedback Methods, respectively. The first part, Evaluation, contains a complete summary of the retrieval results derived from some sixty different text analysis experiments. In each case, significance computations are included which can be used to assess the statistical significance of the reported retrieval results. Conclusions are drawn from the experimental results which may affect the design of future fully-automatic information retrieval systems.

The second and third parts are devoted to a number of experiments in real-time information retrieval, designed to make the user participate in the search process. Specifically, various iterative search strategies are

evaluated based on feedback information returned to the system by the user, following an initial search operation. This feedback information is then used to improve the output obtained during subsequent searches. Evaluation results are presented in part three for many kinds of user feedback strategies.

If the user is relied upon to assist in the search operations, then the system response times must be strictly controlled. In particular, it becomes impossible to perform full searches of the stored document collections which may require the comparison of each stored item with each given search request. Instead, fast searches are required where only small subsections of the available collections are examined. Such "document cluster" searches are examined in part two of this report.

Part 1 on evaluation consists of sections I, II, and III. 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.

Section II by M. E. Lesk contains a write-up of the new evaluation system incorporated into the SMART System for the determination of the statistical significance of the output results. Specifically, for each pair of search and analysis methods being compared, statistical measurements are presented assessing the significance of the differences in performance between the two methods. Thus, it becomes possible to distinguish random variations due to individual request differences from statistically valid results.

Section III by G. Salton and M. E. Lesk is a summary of the SMART evaluation results obtained during 1966 and the first part of 1967 with

collections of 780 document abstracts, 82 short papers, and 200 abstracts in the areas of computer science, documentation, and aerodynamics, respectively. The main (statistically valid) results which are obtained for all collections state that the use of weighted subject identifiers is always more effective than the use of logical identifiers with weights restricted to 0 and 1; the cosine correlation is superior as a method for the comparison of analyzed documents and search requests; furthermore, full document abstracts are far more effective as a source of content identification than document titles alone. A thesaurus process, involving synonym recognition always performs more effectively than methods using the original words, or word stems, included in a document; other dictionary methods, including procedures based on phrase dictionaries and subject hierarchies are not substantially more useful than synonym dictionaries alone. The more effective, fully-automatic text analysis procedures are approximately equivalent in performance to standard retrieval methods based on manually assigned keywords.

The cluster search procedures based on automatically generated document groupings are described in Part 2, consisting of Sections IV to VII. A short description of the computer programs developed both for cluster searching and for the user feedback operations is contained in Section IV by E. Ide, and D. and R. Williamson. The principal search strategies and feedback operations useful in a real-time, user-controlled retrieval environment are then described in Section V by G. Salton, where typical evaluation results are also given.

Several alternative parameters for the generation of document clusters are evaluated in Section VI by R. Grauer and M. Messier.

It is found, in particular, that search systems based on a large number of document groups, each group containing only a few documents, produce generally better results than systems based on fewer clusters of larger size. This same result is also obtained in Section VII by P. C. Leech and R. C. Matlack using document groupings obtained from several kinds of analyzed document representations. Sections VI and VII contain a variety of additional results leading to the generation of more effective document groups useful in partial search systems.

Part 3, consisting of sections VIII to XII, describes the so-called "relevance feedback" process which allows the users to introduce data concerning their information needs in order to obtain improved service. A large variety of feedback strategies is considered in Section VIII by E. Ide. Evaluation results are presented, in particular, for systems in which the user examines a fixed number of documents following each search operation. Situations where a variable number of items is used in each iteration are also treated. It is found that users desiring high-precision should feed back information for only one relevant item in each of two search iterations. Users requiring high recall should, however, examine a larger number of documents after the initial search. Results are also given in Section VIII for a "negative" feedback strategy where information concerning the first nonrelevant item is used to update the search request, in addition to information derived from relevant items.

"Negative" feedback strategies are further studied in Section IX by J. Kelly for the special case where the system fails to produce any useful items for a given customer. In that case, the information to be

fed back is wholly negative in the sense that the available search request is known to be unhelpful, even though it is not clear how it can be improved directly. A negative feedback strategy is described and evaluated which is designed to sweep through the document space in a controlled manner so as to insure that relevant items are eventually picked up if they exist.

A feedback strategy based on a dynamic document space is investigated by S. R. Friedman, J. A. Maceyak, and S. F. Weiss in Section X. This strategy consists in altering the document space by letting both documents and search requests approach the set of relevant items previously identified by the users (in a normal feedback process only the queries are moved but not the documents). It is shown that the system based on document transformations produces greater improvements in search effectiveness than query transformations alone.

Bibliographic information is used as an additional feedback resource in an experimental system described in Section XI by M. Amreich, G. Grissom, and D. Michelson. Specifically, relevant items are identified by the users as before. The user is then allowed to supplement the respective content identifiers by information consisting of the names of authors of related documents, or of bibliographic information cited by the relevant documents. Bibliographic information used as a source of feedback is found to be comparable in effectiveness to feedback information derived from subject identifiers alone. These findings reinforce the importance of previous work on citation tracing and "bibliographic coupling" as an aid in the retrieval process.

In Section XII by H. A. Hall and N. H. Weideman, a new evaluation algorithm is introduced for searches using feedback procedures. Two separate

effects are noted both of which may increase the recall and precision of a feedback search: the shift between successive iterations in the rank of documents which had already been retrieved earlier, called the ranking effect; and the retrieval of new relevant items, called the feedback effect. Evaluation methods are proposed which take into account the feedback effect only, and disregard the ranking effect.

Readers interested in an overview of the work contained in this report should start by looking at sections III, V, VI, and VIII before taking up the remaining studies.