C465

BIBLIOTHEQUE DU CERIST

THE SMART RETRIEVAL SYSTEM

Experiments in Automatic Document Processing

Prentice-Hall
Series in Automatic Computation
George Forsythe, editor

ANSELONE, *Collectively Compact Operator Approximation Theory and Applications to Integral Equations*
ARBIB, *Theories of Abstract Automata*
BATES AND DOUGLAS, *Programming Language/One*, 2nd ed.
BAUMANN, FELICIANO, BAUER, AND SAMELSON, *Introduction to ALGOL*
BLUMENTHAL, *Management Information Systems*
BOBROW AND SCHWARTZ, editors, *Computers and the Policy-Making Community: Applications to International Relations*
BOWLES, editor, *Computers in Humanistic Research*
CRESS, DIRKSEN, AND GRAHAM, *FORTRAN IV with WATFOR and WATFIV*
DANIEL, *The Approximate Minimization of Functionals*
EVANS, WALLACE, AND SUTHERLAND, *Simulation Using Digital Computers*
FIKE, *Computer Evaluation of Mathematical Functions*
FORSYTHE AND MOLER, *Computer Solution of Linear Algebraic Systems*
GAUTHIER AND PONTO, *Designing Systems Programs*
GEAR, *Numerical Initial Value Problems in Ordinary Differential Equations*
GORDON, *System Simulation*
GREENSPAN, *Lectures on the Numerical Solution of Linear, Singular and Nonlinear Differential Equations*
HARTMANIS AND STEARNS, *Algebraic Structure Theory of Sequential Machines*
HEAPS, *An Introduction to Computer Languages*
HULL, *Introduction to Computing*
JOHNSON, *System Structure in Data, Programs, and Computers*
LORIN, *Parallelism in Hardware and Software: An Introduction to Considerations in Real and Apparent Concurrency*
MARTIN, *Design of Real-Time Computer Systems*
MARTIN, *Future Developments in Telecommunications*
MARTIN, *Programming Real-Time Computer Systems*
MARTIN, *Systems Analysis for Data Transmission*
MARTIN, *Telecommunications and the Computer*
MARTIN, *Teleprocessing Network Organization*
MARTIN AND NORMAN, *The Computerized Society*
MATHISON AND WALKER, *Computers and Telecommunications: Issues in Public Policy*
MC KEEMAN, HORNING, AND WORTMAN, *A Compiler Generator*
MINSKY, *Computation: Finite and Infinite Machines*
MOORE, *Interval Analysis*
PYLYSHYN, editor, *Perspectives on the Computer Revolution*
RUSTIN, editor, *Computer Science*
SALTON, *The SMART Retrieval System: Experiments in Automatic Document Processing*
SAMMET, *Programming Languages: History and Fundamentals*
SIMON AND SIKLOSSY, editors, *Representation and Meaning: Experiments with Information Processing Systems*
STERLING AND POLLACK, *Introduction to Statistical Data Processing*
STROUD, *Approximate Calculation of Multiple Integrals*
TAVISS, editor, *The Computer Impact*
TRAUB, *Iterative Methods for the Solution of Equations*
VARGA, *Matrix Iterative Analysis*
VAZSONYI, *Problem Solving by Digital Computers with PL/1 Programming*
WILKINSON, *Rounding Errors in Algebraic Processes*

BIBLIOTHEQUE DU CERIST

C465

GERARD SALTON, *Editor*
Professor of Computer Science
Cornell University

BIBLIOTHEQUE DU CERIST

THE SMART RETRIEVAL SYSTEM

Experiments in Automatic Document Processing

PRENTICE-HALL, INC., Englewood Cliffs, New Jersey

BIBLIOTHEQUE DU CERIST

© 1971 by Prentice-Hall, Inc.
Englewood Cliffs, N.J.

All rights reserved. No part of this book
may be reproduced in any form or by any
means without permission in writing from
the publisher.

Current printing (last digit):
10 9 8 7 6 5 4 3 2 1

13-81425-3
Library of Congress Catalog Card No. 70-159122

Printed in the United States of America

PRENTICE-HALL INTERNATIONAL, INC., *London*
PRENTICE-HALL OF AUSTRALIA, PTY. LTD., *Sidney*
PRENTICE-HALL OF CANADA, LTD., *Toronto*
PRENTICE-HALL OF INDIA PRIVATE LIMITED, *New Delhi*
PRENTICE-HALL OF JAPAN, INC., *Tokyo*

BIBLIOTHEQUE DU CERIST

PREFACE

The automatic SMART document retrieval system was designed at Harvard University between 1961 and 1964, and has been operating on IBM 7094 and 360 equipment both at Harvard and at Cornell University for several years. The system takes documents and search requests in the natural language, performs a fully automatic content analysis of the texts using one of several dozen programmed language analysis methods, matches analyzed documents with analyzed search requests, and retrieves for the user's attention those stored items believed to be most similar to the submitted queries.

Unlike the other computer-based retrieval systems, the SMART system does not rely on manually assigned keywords or index terms for the identification of documents and search requests, nor does it use primarily the frequency of occurrence of certain words or phrases included in the texts of documents. Instead, an attempt is made to go beyond simple word-matching procedures by using a variety of intellectual aids in the form of synonym dictionaries, hierarchical arrangements of subject identifiers, statistical and syntactic phrase generation methods, and the like, in order to obtain content identifications useful for the retrieval process.

By comparing the retrieval performance obtained with the various programmed procedures, the SMART system can be used as a unique experimental tool for the evaluation in a controlled laboratory environment of many fully automatic language analysis methods. In addition, the system

has been used to simulate a user-environment by making it possible for the user to participate in the search process. Specifically, the system utilizes feedback information supplied by the user during the search to construct improved search formulations, and to generate document representations reflecting the interests of the user population. By combining automatic text processing methods with interactive search and retrieval techniques, the SMART system may then lead to the design and implementation of modern information services of the type which may become current in operational environments some years hence.

Most of the documentation pertaining to the SMART system design and to the experimental results obtained with the system over the last few years is contained in a set of book-size scientific reports entitled "Information Storage and Retrieval," known in brief as "the ISR reports." The first ten of these (ISR-1 to ISR-10) were issued at Harvard between November 1961 and March 1966, and the last seven at Cornell. At the time of this writing, the most recent volume issued is ISR-17, dated September 1969.

The ISR reports covering the SMART system are not generally available in the open market; moreover, the information contained in the reports is difficult to assimilate, being dispersed over a large number of volumes including many thousands of pages. For this reason it has seemed advisable to collect in an organized manner, as a single book, the most important contributions contained in the earlier reports.

The present volume thus consists of updated versions of twenty-seven studies taken from the material contained in the ten most recent scientific reports (ISR-8 to ISR-17). Among the material covered are theoretical developments, including the derivation of system evaluation measures, language analysis techniques, document grouping techniques, and adaptive space transformation methods, as well as experimental studies relating to the evaluation of document analysis methods, interactive user feedback procedures, partial document searches based on clustered file organizations, and comparisons between the SMART system and more conventional operational information systems.

The present text is organized into eight major parts, entitled, respectively, The SMART System, Evaluation Viewpoint and Parameters, Language Analysis, Cluster Generation and Search, Basic Feedback Runs, Feedback Refinements, Document Space Transformation, and Operational Comparisons. Each part can be read independently of the remainder, and none of the material requires more than an elementary knowledge of mathematics or computer programming.

The text should be of greatest value as a reference volume for the professional practitioner interested in the design and operations of automatic information systems. It should also be useful as a text of readings in the area of automatic information retrieval for the more mature students enrolled in courses in applied mathematics, computer and information science, or library science. In this latter role, the studies contained in this volume might serve as a point of departure for term projects and for experimental work in modern information processing.

Part I of this volume, consisting of Chapters 1 and 2, covers the basic design of the SMART system. Chapter 1 by G. Salton deals with the background relating to the SMART project and contains a summary of evaluation results already obtained and a report of future plans. Chapter 2 by D. Williamson, R. Williamson, and

BIBLIOTHEQUE DU CERIST

M. E. Lesk is a description of the implementation of the SMART system as it is presently operating at Cornell on an IBM 360/65.

The main systems evaluation parameters are covered in Part II, consisting of Chapters 3 to 5. Chapters 3 and 4 by J. J. Rocchio, Jr., contain a derivation of the normalized recall and precision evaluation measures and of the procedures used to obtain average retrieval results valid over many information searches. Chapter 5 by E. M. Keen is a detailed description of the design of a retrieval evaluation system.

The language analysis problems and relevant evaluation results are contained in Part III, consisting of Chapters 6 to 9. Chapter 6 by G. Salton and M. E. Lesk deals with the general problem of language analysis and dictionary construction useful in a content analysis system. Chapter 7 also by G. Salton and M. E. Lesk is a summarization of evaluation results obtained with SMART by processing document collections in the areas of aerodynamics, computer engineering, and documentation. A thorough analysis of the performance of user queries in the area of documentation is contained in Chapter 8 by E. M. Keen. The final chapter of Part III, number 9, by G. Salton covers an extension of the language analysis procedures originally implemented for English items to a collection of German language documents.

Part IV, consisting of Chapters 10 to 13, contains a description of automatic document classification procedures, and of the partial search methods based on document clusters. Chapter 10 by G. Salton is a basic description of cluster searching. Chapter 11 by R. T. Grauer and M. Messier covers the evaluation of a clustering process due to Rocchio. Additional, more efficient methods for automatic document classification are described in Chapter 12 by R. T. Dattola. Finally, Chapter 13 by S. Worona introduces an information search process based on the generation of request, rather than document, clusters.

The standard query transformation process based on user feedback is covered in Part V, consisting of Chapters 14 to 17. The main theoretical considerations relating to the construction of optimal user queries by relevance feedback methods are contained in Chapter 14 by J. J. Rocchio, Jr., and the principal feedback evaluation methods are described in Chapter 15 by G. Salton. A thorough analysis of feedback retrieval is included in Chapter 16 by E. Ide, and the last chapter, number 17, by C. Cirillo, Y. K. Chang, and J. Razon examines a number of novel feedback evaluation methods which circumvent some of the difficulties of the standard feedback evaluation methodology.

Certain refinements of the interactive search process are covered in Part VI of this volume, consisting of Chapters 18 to 22. The main distinctions between positive and negative feedback procedures are described in Chapter 18 by E. Ide and G. Salton; this chapter also contains a discussion of the relevance feedback process in the environment of a clustered information file. A query-splitting procedure, useful when nonhomogeneous sets of retrieved documents are identified as relevant during the search process is evaluated in Chapter 19 by A. Borodin, L. Kerr, and F. Lewis. A novel, experimental procedure for the implementation of negative feedback is contained in Chapter 20 by J. Kelly. Chapter 21 by J. S. Brown and P. D. Reilly deals with refined query modification methods in which the document and query terms receive individual treatment depending on the particular term characteristics. Finally Chapter 22 by D. Michelson, M. Amreich, G. Grissom, and E. Ide describes

BIBLIOTHEQUE DU CERIST

procedures incorporating author information and bibliographic references into the feedback process.

The use of document, rather than query, transformations, made possible in most interactive retrieval environments is described in Part VII, consisting of Chapters 23 and 24. A procedure for document space transformation is introduced in Chapter 23 by S. R. Friedman, J. A. Maceyak and S. F. Weiss; a brief evaluation of the suggested procedure is also given. A modification of the transformation process covered in Chapter 23 is examined in detail in Chapter 24 by T. L. Brauen, and a thorough evaluation is presented both for the positive and the negative space modification processes.

The last part, number VIII, consists of Chapters 25 to 27, and covers various aspects connected with the operational evaluation of the SMART system. The use of a large variety of information displays of the kind presently available in an interactive retrieval environment with graphic or typewriter console equipment is described and evaluated in Chapter 25 by M. E. Lesk and G. Salton. Chapter 26 by M. E. Lesk and G. Salton covers an experiment designed to determine the importance of user relevance assessments for retrieval evaluation. Finally, in Chapter 27 by G. Salton a preliminary comparison is made between the retrieval effectiveness of the fully automatic SMART system, and the well-known MEDLARS system operating at the National Library of Medicine in Washington.

Readers wishing a quick overall view might restrict their attention to Parts I and III, and possibly to the operational problems described in Part VIII. The problems of content analysis are covered in detail in parts I, II, III and also VIII, whereas the interactive search procedures are contained in Parts IV, V, VI and VII. Readers interested in the problems of automatic language analysis should thus concentrate on the first set of chapters, while persons concerned with problems of file organization and file search should study the other half (Parts IV to VII).

The SMART system consisting of dozens of programmed routines and several hundred thousand machine instructions could not have been implemented without the help of the many programmers, analysts, and students who over the years have participated in the project, both at Harvard and at Cornell. Among those particularly instrumental who have extensively contributed to the design and programming phases are Dr. E. H. Sussenguth, Jr., now at IBM Corporation, Dr. J. J. Rocchio, Jr., now with International Computing Company, Dr. M. E. Lesk, now at Bell Telephone Laboratories; Mr. E. M. Keen, now at the College of Librarianship in Aberystwyth, Wales; and Mr. Robert E. Williamson, a graduate student at Cornell University. A number of other present or former students have performed important evaluation studies, including, in particular, T. Brauen, R. Dattola, and E. Ide.

To all these individuals, and to past and future users of the system, I am indebted for help and advice, and for extraordinary patience in bearing with the imperfections of an experimental system. I also wish to thank the McGraw-Hill Book Company for permission to include certain tabular material in Chapters 6 and 7. Finally, I am particularly grateful to the National Science Foundation whose continuous support over many years made it possible to design the system and to perform the research leading to the present implementation of the SMART system.

Ithaca, New York GERARD SALTON

BIBLIOTHEQUE DU CERIST

BIBLIOTHEQUE DU CERIST

CONTENTS

BIBLIOTHEQUE DU CERIST

BIBLIOTHEQUE DU CERIST

BIBLIOTHEQUE DU CERIST

BIBLIOTHEQUE DU CERIST

BIBLIOTHEQUE DU CERIST

BIBLIOTHEQUE DU CERIST

BIBLIOTHEQUE DU CERIST

BIBLIOTHEQUE DU CERIST