THE COMPUTATION LABORATORY

Harvard University

Cambridge, Massachusetts

Scientific Report No. ISR-10

INFORMATION STORAGE AND RETRIEVAL

to

The National Science Foundation

Cambridge, Massachusetts
March 1966

Gerard Salton
Project Director

DOCUMENT RETRIEVAL SYSTEMS - OPTIMIZATION AND EVALUATION

A thesis presented

by

Joseph John Rocchio, Jr.

to

The Division of Engineering and Applied Physics

in partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

in the subject of

Applied Mathematics

Harvard University

Cambridge, Massachusetts

March 1966

PREFACE

This thesis presents a functional model for document retrieval systems. Based on this model, a search request optimization algorithm and an automatic classification algorithm applicable to efficient storage organization are derived. In addition some novel evaluation measures for document retrieval systems are proposed.

The writer is deeply indebted to Professor Gerard Salton for his guidance during the conception and development of this thesis, and for his confidence, encouragement, and continuing support.

Particular thanks are also due to Mr. Michael Lesk for his valuable assistance with respect to programming problems and with operational aspects of the SMART system, and to the other members of the SMART project at the Harvard Computation Laboratory. In addition the editorial assistance and moral support of Mary Rocchio, the author's wife, were of immeasurable value.

Finally the writer wishes to acknowledge the support of this research by the National Science Foundation under grant GN 360, and by the Bell Telephone Laboratories under contract with Harvard University.

TABLE OF CONTENTS

TABLE OF CONTENTS (continued)

TABLE OF CONTENTS (continued)

LIST OF FIGURES

LIST OF FIGURES (continued)

LIST OF TABLES

SYNOPSIS

A model for document retrieval systems consisting of the functional elements: indexing, search request formulation, and request-document matching is examined in this thesis. Analysis of the request formulation function leads to the definition of an optimality criterion and to a request optimization algorithm suitable for use in a system environment which allows iterative searching and real time user-system interaction. The optimality criterion has applicability to the evaluation of index languages, and generally to the design of evaluation tests for document retrieval systems. Investigation of the request-document matching leads to a novel automatic classification algorithm applicable to metric comparison measures, and useful for establishing an efficient storage organization. Finally the statistical basis for the evaluation of document retrieval systems is reviewed and some novel performance measures are proposed which are particularly suited to systems which induce a retrieval ordering on the members of the searched collection.

Chapter 1 is introductory in nature and attempts to define the area of discourse and its relation to the general field of information retrieval. A general model for document retrieval systems is introduced, and the basic functional elements of this model are briefly outlined. This material draws heavily on the work of Salton and the SMART automatic document retrieval project, as well as the general literature of the field.

The indexing component of the model is discussed in chapter 2 which is primarily descriptive in nature. The work of a number of researchers in this area is cited in describing the development and current trends of automatic content analysis. Particular emphasis is placed on concept vector indexing techniques which incorporate thesaurus type semantic associations. The SMART system generates document and query index images by such techniques and these were used in obtaining the experimental results presented in Chapters 3 and 4. The original contributions in chapter 2 are related to the techniques proposed for index image optimization.

A search request optimization algorithm analytically derived from an assumed optimality criterion is presented in Chapter 3. The optimization algorithm and the notion of request optimization by an iterative sequence of retrieval operations are original with the author. In addition the notion of testing index language devices by the use of optimal search requests is original. Experimental results illustrating the optimization process are presented. These were derived by a simulation which was coded and run on the IBM 7094 in conjunction with the SMART retrieval system. A search request formulation based on this optimization technique offers the promise of improving the performance of document retrieval systems, given the current state of development of computer time sharing and man-machine communication technology.

Chapter 4 presents an original automatic document classification algorithm, heuristically motivated by considerations of search efficiency and by the functional nature of query-document matching operations. This algorithm was coded in Fortran and run on

the IBM 7094 producing several classifications of a collection of 405 document index images (incorporated into the SMART system). Some experimental results illustrating the use of the classifications for improving search efficiency are presented. Document classification of the type described is novel, in that it is proposed as a direct adjunct to the query-document matching operation. Normally, automatic classification algorithms have been considered as replacements for manual classification or indexing.

The statistical basis for the evaluation of document retrieval systems is discussed in Chapter 5. Several of the topics considered are based on previous work which is cited by bibliographic reference. The organization and presentation as well as some of the conclusions drawn are original. In addition some novel performance statistics are derived which are particularly applicable to query-document matching operations possessing a high degree of discrimination such as the correlation measure of the assumed model. Each of the statistics derived is capable of describing overall system performance with a single parameter in contrast with several of the evaluation measures in current use.

DOCUMENT RETRIEVAL SYSTEMS - OPTIMIZATION AND EVALUATION