INTRODUCTION TO MODERN INFORMATION RETRIEVAL

COMPUTER
SCIENCE
SERIES

Introduction to Modern Information Retrieval

McGraw-Hill Computer Science Series

Ahuja: Design and Analysis of Computer Communication Networks
Allen: Anatomy of LISP
Barbacci and Siewiorek: The Design and Analysis of Instruction Set Processors
Bell and Newell: Computer Structures: Readings and Examples
Donovan: Systems Programming
Gear: Computer Organization and Programming
Givone: Introduction to Switching Circuit Theory
Goodman and Hedetniemi: Introduction to the Design and Analysis of Algorithms
Hamacher, Vranesic, and Zaky: Computer Organization
Hamming: Introduction to Applied Numerical Analysis
Hayes: Computer Architecture and Organization
Hellerman: Digital Computer System Principles
Hellerman and Conroy: Computer System Performance
Katzan: Microprogramming Primer
Keller: A First Course in Computer Programming Using PASCAL
Liu: Elements of Discrete Mathematics
Liu: Introduction to Combinatorial Mathematics
MacEwen: Introduction to Computer Systems: Using the PDP-11 and Pascal
Madnick and Donovan: Operating Systems
Manna: Mathematical Theory of Computation
Newman and Sproull: Principles of Interactive Computer Graphics
Nilsson: Problem-Solving Methods in Artificial Intelligence
Payne: Introduction to Simulation: Programming Techniques and Methods of Analysis
Rice: Matrix Computations and Mathematical Software
Salton and McGill: Introduction to Modern Information Retrieval
Shooman: Software Engineering: Design, Reliability, and Management
Siewiorek, Bell, and Newell: Computer Structures: Principles and Examples
Stone: Introduction to Computer Organization and Data Structures
Stone and Siewiorek: Introduction to Computer Organization and Data Structures: PDP-11 Edition
Tonge and Feldman: Computing: An Introduction to Procedures and Procedure-Followers
Tremblay and Bunt: An Introduction to Computer Science: An Algorithmic Approach
Tremblay and Bunt: An Introduction to Computer Science: An Algorithmic Approach, Short Edition
Tremblay and Manohar: Discrete Mathematical Structures with Applications to Computer Science
Tremblay and Sorenson: An Introduction to Data Structures with Applications
Tucker: Programming Languages
Wiederhold: Database Design

McGraw-Hill Advanced Computer Science Series

Davis and Lenat: Knowledge-Based Systems in Artificial Intelligence
Kogge: The Architecture of Pipelined Computers
Lindsay, Buchanan, Feigenbaum, and Lederberg: Applications of Artificial Intelligence for Organic Chemistry: The Dendral Project
Nilsson: Problem-Solving Methods in Artificial Intelligence
Wulf, Levin, and Harbison: Hydra/C.mmp: An Experimental Computer System

Introduction to Modern Information Retrieval

Gerard Salton
Professor of Computer Science
Cornell University

Michael J. McGill
Associate Professor of Information Studies
Syracuse University

McGraw-Hill Book Company

New York St. Louis San Francisco Auckland Bogotá Hamburg
Johannesburg London Madrid Mexico Montreal New Delhi
Panama Paris São Paulo Singapore Sydney Tokyo Toronto

This book was set in Times Roman by Progressive Typographers.
The editors were Charles E. Stewart and James E. Vastyan;
the production supervisor was John Mancia.
The drawings were done by VIP Graphics.
R. R. Donnelley & Sons Company was printer and binder.

INTRODUCTION TO MODERN INFORMATION RETRIEVAL

1 2 3 4 5 6 7 8 9 0 DODO 8 9 8 7 6 5 4 3 2

ISBN 0-07-054484-0

Library of Congress Cataloging in Publication Data

Salton, Gerard.
Introduction to modern information retrieval.

(McGraw-Hill computer science series)
Includes indexes.
1. Information storage and retrieval systems.
I. McGill, Michael J. II. Title. III. Series.
Z699.S313 025'.04 81-20843
ISBN 0-07-054484-0 AACR2

Contents

Preface

An information retrieval system is an information system, that is, a system used to store items of information that need to be processed, searched, retrieved, and disseminated to various user populations. Information retrieval systems thus share many of the concerns of other information systems, such as data base management and decision support systems. In particular, it is necessary to choose efficient organizations for the stored records, rapid search procedures capable of finding items of interest in specific cases, and effective methods for disseminating the retrieved data and interacting with the system users.

Information retrieval systems are normally used to handle bibliographic records and textual data. This is in contrast to data base management and management information systems that process structured data, and to question-answering systems that use complex information organizations and inference procedures designed to answer questions in particular subject areas. In an extended sense, however, any information system designed to augment the state of human knowledge and to aid human activities does utilize concepts and procedures from information storage and retrieval.

Today, information processing activities are carried out with the assistance of automatic equipment. Thus, a direct link exists between information retrieval and computer science. On the other hand, information retrieval also takes on aspects of behavioral science, since retrieval systems are designed to aid human activities.

Most practitioners interested in the design and operations of actual retrieval systems are concerned only about applied computer science. However, many topics in theoretical computer science are also of direct importance to information retrieval, including, for example, information theory, probability theory, computational semantics, and programming theory and algebra. Techniques from these disciplines may be used to build information retrieval models and to obtain insights into various aspects of retrieval theory and practice.

Although information retrieval is mentioned in many computer science curriculum proposals, retrieval courses are often replaced in practice by material on data structures and data base systems where attractive approaches have been developed for formalization and abstraction. The study of information retrieval is thus frequently carried out in library science, information science, and information management schools. In these environments, the mathematical foundations necessary to make a substantial impact are often omitted from retrieval system courses.

This text is aimed at increasing the understanding of modern information retrieval by students of computer science as well as by students of information science and management science. The book covers the basic aspects of information retrieval theory and practice, and also relates the various techniques to the design and evaluation of complete retrieval systems. The book is introductory in the sense that no prior knowledge of retrieval methodology is assumed; it is modern because currently active trends and developments are examined. The text concentrates in particular on the description of the concepts, functions, and processes of interest in retrieval rather than on the detailed operations of any one existing retrieval system. In order to keep the material at an introductory level, the more advanced mathematical aspects of retrieval theory have been deemphasized or simplified. The text should thus be accessible to students with only a cursory knowledge of the operations of digital computers and only a superficial exposure to computer programming. More advanced readers can supply the relevant mathematical background by consulting the references given.

The text begins with an introduction and a description of the main retrieval processes incorporated into existing, operational systems based on keyword indexing and Boolean query formulations (Chapters 1 and 2). Chapter 3 contains a detailed explanation of modern automatic indexing techniques with evaluation results and assessments of the importance and practical usefulness of the techniques. Experimental retrieval systems, based in part on fully automatic analysis, search, and retrieval methods, are covered in Chapter 4 with emphasis on the design of the SMART and SIRE systems developed by the authors. The main evaluation techniques used to assess the effectiveness and efficiency of information retrieval systems are covered in Chapter 5 with emphasis on the use of the well-known recall and precision measures. Chapter 6 deals with important techniques usable in the design of future systems, such as automatic classification methods, query negotiation, and reformulation processes used in on-line environments, collection restructuring, and bibliographic citation processing. Language processing methods useful in retrieval, including current syntactic and semantic methodologies, and artificial intelligence approaches to language understanding are examined in Chapter 7. Chapter 8 introduces specialized hardware useful in retrieval, such as parallel processing devices, array processors, and special back-end search devices useful for manipulating and searching large data bases. Also covered are modern techniques used for dictionary searching and for automatic text scanning systems. The relationship between information retrieval and other information systems, such as data base and decision support systems, is examined in detail in Chapter 9, and the current work in data base processing and data base management is described. Finally the expected future directions and developments in information retrieval are covered in Chapter 10, including the importance and likely effect of personal computers, word processing, advanced display systems, and paperless information systems.

The text should be useful for computer science as well as library science and information science students on a junior-senior level in college or for beginning graduate students. The book can also serve the professional reader as an introduction to the design and operations of information retrieval and management information systems. A common core for computer science as well as information science readers is contained in Chapters 1, 3, 4, and 10. Computer science audiences should profit in addition from a complete treatment of Chapters 5, 6, and 8, since these chapters emphasize the more mathematical aspects of the field and the connection to software and hardware implementations. Library and information science readers, on the other hand, should concentrate on Chapters 2, 7, and 9 in addition to the core to gain a thorough understanding of conventional retrieval and other information systems and of the language analysis methods useful for processing natural language texts.

To simplify the reader's task, the material has been graded for technical difficulty:

*Sections marked with a single asterisk contain technical material that may be difficult for some readers. Often a modest computer science background may be useful, or an acquaintance with elementary algebra or basic probability theory. This material is considered important, and the reader is encouraged to read the section and obtain an understanding of the content.

**Sections identified by two asterisks contain technical material at a somewhat more detailed level. A particular procedure may be covered in detail; alternatively, a theory may be introduced requiring some technical know-how. Readers who find this material difficult may wish to skim the section rather than dwell on the details.

Sections not marked by * or ** should be accessible to all readers without special background.

The following sample curricula will provide complete coverage of the principal aspects of retrieval system design and operations:

	Computer science	Information science and management science
Chapter 1 (Core)	What is information retrieval? Functional view of retrieval	What is information retrieval? Functional view of retrieval
Chapter 2 (IS emphasis)	Basic set theory inherent in list processing	Standard Boolean operations Standard retrieval Conventional systems
Chapter 3 (Core)	Theory of automatic indexing Term weighting and associative indexing Basic evaluation results	Theory of automatic indexing Term weighting and associative indexing Basic evaluation results
Chapter 4 (Core)	The SMART system Relevance feedback and cluster searching Weighted retrieval in Boolean systems (SIRE)	The SMART system Relevance feedback and cluster searching Weighted retrieval in Boolean systems (SIRE)
Chapter 5 (CS emphasis)	Mathematics of evaluation Evaluation parameters and computational aspects	Basic definition of recall-precision parameters Cost evaluation
Chapter 6 (CS emphasis)	Term relevance theory Cluster generation and search Pseudoclassification Document space alteration and dynamic file processing	Use of citations in information search systems
Chapter 7 (IS emphasis)	ATN grammars Criterion tree processing Concept representation	Language analysis Syntax and semantics Context-free and context-sensitive grammars Concept representation in information systems
Chapter 8 (CS emphasis)	Basic hardware devices Parallel processing techniques Microprocessors Dictionary search methods String search algorithms	Basic hardware and parallel processing techniques Microprocessors Text scanning machines
Chapter 9 (CS and IS emphasis)	Relationship of information retrieval to other information systems Data base systems and models	Relationship of information retrieval to other information systems File processing, accessing, searching File security, data structures
Chapter 10 (Core)	Future directions Text input Distributed architecture New retrieval theories Advanced information systems	Future directions Text input Distributed architecture Mixed information retrieval systems Paperless information systems

By limiting the coverage to the more basic aspects of the various topics, the material can be assimilated in a one-semester course. A second semester may be required if the various algorithms and techniques—for word stem generation, pseudoclassification, string searching, and so on—are covered in detail, and if additional sources are consulted.

A number of graduate students at Cornell and Syracuse universities have made substantial contributions to the design of the SMART and SIRE systems, including, in particular: Robert E. Williamson, Clement T. Yu, Chung Shu Yang, Anita Wong, Harry Wu, and Edward A. Fox at Cornell and Terry Noreault, Jennifer Kuehn, Judy Tessier, and Matthew Koll at Syracuse. Several readers have reviewed the manuscript and made many valuable comments and suggestions, including, in particular: Professor Richard H. Austing of the University of Maryland, Professor Michael D. Cooper of the University of California at Berkeley, Professor Jeffrey Katzer of Syracuse University, Dr. Michael E. Lesk of Bell Laboratories, Professor J. F. Nunamaker of the University of Arizona, and Professor Linda C. Smith of the University of Illinois. The authors have also profited from discussion with many colleagues and friends. Some early material was typed by Peggy Montgomery at Syracuse. Geri Pinkham at Cornell has typed the complete manuscript over several times with unusual speed and competence; in the process she became familiar with the intricacies of an automated text editing system. Without her help the text would have remained in manuscript form for a long time to come. Edward Fox and Elena Seifrid have also helped to produce a version ready for automatic typesetting. The writers are greatly indebted to all these individuals for guidance and assistance.

Gerard Salton
Michael J. McGill