

Automatic Text Processing

The Transformation, Analysis, and
Retrieval of Information by Computer

Gerard Salton

Cornell University



ADDISON-WESLEY PUBLISHING COMPANY

Reading, Massachusetts • Menlo Park, California
New York • Don Mills, Ontario • Wokingham, England
Amsterdam • Bonn • Sydney • Singapore
Tokyo • Madrid • San Juan

This book is in the Addison-Wesley Series in Computer Science

Michael A. Harrison, Consulting Editor

Many of the designations used by manufacturers and sellers to distinguish their products are claimed as trademarks. Where those designations appear in this book, and Addison-Wesley was aware of a trademark claim, the designations have been printed in initial caps or all caps.

The programs and applications presented in this book have been included for their instructional value. They have been tested with care, but are not guaranteed for any particular purpose. The publisher does not offer any warranties or representations, nor does it accept any liabilities with respect to the programs or applications.

Library of Congress Cataloging-in-Publication Data

Salton, Gerard.

Automatic text processing: the transformation, analysis, and retrieval of information by computer / by Gerard Salton.

p. cm.

Bibliography: p.

Includes index.

ISBN 0-201-12227-8

1. Text processing (Computer science) I. Title.

QA76.9.T48S25 1989

005—dc19

88-467

CIP

Reprinted with corrections December, 1988

Copyright © 1989 by Addison-Wesley Publishing Company, Inc. All rights reserved. No part of this publication may be reproduced, stored in a retrieval system, or transmitted, in any form or by any means, electronic, mechanical, or photocopying, recording, or otherwise, without the prior written permission of the publisher. Printed in the United States of America. Published simultaneously in Canada.

4 5 6 7 8 9 10 HA 959493

Preface

The current period is known as the information age because more information is generated about more topics than ever before. In this complex world, relevant information is often needed to carry out the tasks at hand and to make intelligent decisions. When large data banks of information are collected and stored, it is difficult to find the data actually needed at a given time, and to distinguish relevant from extraneous data. For this reason, electronic search aids are widely used to process, store, and retrieve information items on demand.

The information of interest at any particular time takes various forms. In particular, standard written data and natural-language texts must be distinguished from spoken utterances and speech sounds, and from graphical and pictorial data. Textual information is primarily important because text is used universally to convey information and to communicate. In addition, text can be automatically processed more easily and less expensively than either speech or pictures.

This book deals with the whole area of automatic text-processing — that is, the handling of texts using automatic equipment. The aim is not to teach laymen or humanists how to program computers to manipulate text, nor to teach scientists language-processing skills. In-

stead, the book examines the area of text processing as a whole, describing various text-processing methodologies and identifying those tasks now undertaken routinely, while also discussing more experimental procedures not yet ready for operation. For example, it is conceptually easy to take the English text which makes up this book and determine the number of occurrences in the text of the word "information." It is more difficult to identify all the sections in the book that deal with "information storage and retrieval," in part because the words "information" and "retrieval" do not occur explicitly in some relevant sections. It is even more difficult to find the sections exhibiting stylistic similarities with the style used in this preface. Indeed, such a request cannot be processed without specifying the perceived stylistic features characterizing the preface. Analogously, it is very difficult to devise effective methods for retrieving from a library all books whose opinions about the mechanization of text processing reflect the opinions expressed here.

This should be a useful reference for users of text-processing systems and designers of text-processing routines. It can also serve as a textbook in programs of computer science and engineering, library and information science, computational linguistics, as well as programs about relations among science, technology, and society. Various parts of the text have been used in a text-processing course taught at Cornell University to upper-level computer-science undergraduates and first-year graduate students.

The book is divided into four main parts. The introduction, Chapters 1, 2, and 3, covers the existing computer environment and the automated office situation, in which text processing is of particular interest. The second part, Chapters 4 to 7, covers the main word-processing areas, which treat texts on the level of individual words. This includes text editing and formatting, properly termed "word processing" in the standard literature. Also included in Part 2 are text-compression methods designed to reduce the size of stored texts, text encryption methods designed to hide the meaning of the texts, and file-accessing methods used to access and search mechanized text files.

Part 3, Chapters 8, 9, and 10, covers text-retrieval systems whose operations are normally based on text units larger than single, individual word forms. Included is an examination of conventional text-retrieval systems based on automatic text scanning as well as conventional indexed text searches. Simple text analysis, and so-called automatic indexing systems designed to assign content identifiers to texts, are also described. Finally, advanced text-retrieval systems are considered that may be based on automatic text classification and complex Boolean query formulations.

Part 4, Chapters 11, 12, and 13, covers the main language-analysis and language-processing topics in which text meaning and text under-

standing are of principal concern: syntactic and semantic language-analysis methods that determine language structure and text content, and modern knowledge-based text processing. Various applications of linguistic procedures are also described including automatic text extracting and abstracting, text generation, and text translation. The book ends with an examination of paperless information systems that process speech and graphics information as well as text. Various electronic-information systems are covered such as electronic mail and message systems, automatic publication systems, and electronic books and libraries.

Each chapter can be read independently, but not every chapter will be equally accessible to every reader. In particular, the more mathematical treatment of text compression and encryption in Chapters 5 and 6 and some of the advanced retrieval methods of Chapter 10 are intended for those with technical training. Specialized sections or subsections, or those that require a mathematical background, are appropriately marked.

The following chapter arrangement can be used for a one-semester course for upper-level undergraduate and beginning graduate students in various disciplines (see also the figures on pages vi and vii):

Computer science and related subjects	Part 2 (Chapters 4–7) on compression, encryption and file access; Part 3 (Chapters 8–10) on automatic information retrieval
Linguistics and language processing	Part 1 (Chapters 1–3); Part 3 (Chapter 9, on document analysis), Part 4 (Chapters 11–13) on language processing
Library and information science, science and technology programs	Part 1 (Chapters 1–3); Chapter 4 of Part 2 on word processing; Part 3 (Chapters 8 and 9) on conventional retrieval; Part 4 (Chapters 11 and 13) on language analysis and paperless information systems

Part 1. The Information-Processing Environment

	Trained computer-science students and computer professionals	Computational linguists and language-processing researchers	Students in library and information science or science and technology programs
Chapter 1: An introduction to various types of information, including text, speech, and graphics		✓	✓
Chapter 2: Modern computer architecture and various computing devices and their capabilities		✓	✓
		may skip details of input-output and peripheral equipment (Sections 2.3.1, 2.3.2, 2.3.3)	
Chapter 3: An introduction to office automation and file-management systems		✓	✓
		may skip relational database management system (Sections 3.3.2, 3.3.3, 3.3.4)	
go to Chapter 4			

Part 2. Word Processing and File Access

	Trained computer-science students and computer professionals	Computational linguists and language-processing researchers	Students in library and information science or science and technology programs
Chapter 4: An introduction to text editing and formatting and to automatic typesetting	may be skipped by persons experienced in text editing	✓	✓
Chapter 5: Statistical language analysis and basic text-compression methods	✓	Section 5.1 may be covered	skip except for Section 5.1
Chapter 6: Introduction to text encryption and a review of basic text-encryption methods	✓		
Chapter 7: File-access methods for single-key and multiple-key search statements	those familiar with data structures may start with Section 7.8		
		go to Chapter 9	go to Chapter 8

Part 3. Information Retrieval Systems

	Trained computer-science students and computer professionals	Computational linguists and language-processing researchers	Students in library and information science or science and technology programs
Chapter 8: A review of conventional text-retrieval and text-scanning methods	✓ for fast reading skip Section 8.5	Go to Chapter 9 ↓	✓ for fast reading skip Section 8.5
Chapter 9: Automatic text-indexing and content-analysis methods	✓	Section 9.3.3 may be skipped ↓	Section 9.3.3 may be skipped ↓
Chapter 10: Advanced retrieval system based on abstract retrieval models, automatic classification, and advanced interfaces	✓	go to Chapter 11	go to Chapter 11

Part 4. Text Analysis and Language Processing

	Trained computer-science students and computer professionals	Computational linguists and language-processing researchers	Students in library and information science or science and technology programs
Chapter 11: A review of approaches to the content analysis of written text, including syntactic, semantic, and knowledge-based approaches	↓	✓	Section 11.5 may be skipped ✓
Chapter 12: Applications of linguistic text processing, including spelling error detection and correction, abstracting, text generation, and translation	↓	Section 12.4 may be skipped ✓	Section 12.4 may be skipped ✓
Chapter 13: Processing of electronic text, speech, and graphics and applications such as teleconferencing, mail and message handling, and electronic publishing	optional depending on interest ↓	✓	✓

Acknowledgements

I am deeply indebted to a number of individuals who have made valuable suggestions about the content and form of this material, especially Dr. Michael E. Lesk of Bell Communications Research, Professor Michael A. Harrison of the University of California at Berkeley, Dr. Michael Lebowitz of Morgan Stanley and Company, Dr. Richard J. Beach of Xerox Palo Alto Research Center, and a number of former and present graduate students in information retrieval at Cornell, including Dr. Edward A. Fox, Dr. Ellen Voorhees, Dr. Joel Fagan, and Christopher Buckley. I am also most grateful to Professor Giovanni Coray of the Mathematics Department at the Ecole Polytechnique Federale in Lausanne (EPFL) and to Professor Charles Stuart, the chairman of that department, for having made available the exceptionally nice surroundings of the EPFL during a sabbatical year in Lausanne in 1984-85 when I wrote the first chapters of this book. Special thanks are due to my colleagues at Cornell, especially the two recent chairmen of the Computer Science Department, Professor David Gries and Professor John Hopcroft, for giving me the time to write a book when many other departmental matters might have taken precedence. Finally, I am deeply grateful to Geri Pinkham for once again spending much time to prepare this manuscript on the departmental word-processing equipment with her usual diligence and competence, and to Margaret Schimizzi and Teresa Leidenfrost, who helped draw the figures and prepare the typescript. I thank all these individuals for their guidance and assistance.

Ithaca, New York

Gerard Salton

Contents

Part 1: The Information-Processing Environment 1

The Information Environment 3

- 1.1 Automatic Information Processing 3
- 1.2 Types of Information 4
 - 1.2.1 Text Processing 5
 - 1.2.2 Speech Processing 6
 - 1.2.3 Graphics Processing 7
 - 1.2.4 Semantic and Behavioral Processing 8

2 The Computer Environment 12

- 2.1 Computer Architecture 12
 - 2.1.1 Large versus Small Machines 12
 - 2.1.2 Sequential versus Parallel Computing 13
 - 2.1.3 Multiprocessor and Multicomputer Configurations 16
 - 2.1.4 Types of Computers 18
- 2.2 Storage Technology 19
- 2.3 Input-Output and Peripheral Equipment 23
 - ◆ 2.3.1 Terminal Equipment 23
 - ◆ 2.3.2 Printing Equipment 25
 - ◆ 2.3.3 Document Input 29
- 2.4 Computer Networks 31
- 2.5 Integrated Computing Systems 36

3	The Automated Office	41
3.1	The Office Environment	41
3.2	Analyzing Office Systems	43
3.3	File-management Systems	47
3.3.1	System Characteristics	47
◆ 3.3.2	Relational Database Systems	50
◆ 3.3.3	Relational Data Manipulations	54
◆ 3.3.4	Data Security, Integrity, and Recovery	60
3.4	Office Display Systems	61
3.5	Office-Information Retrieval	63

Part 2: Word Processing and File Access 71

4	Text Editing and Formatting	73
4.1	Introduction	73
4.2	Approaches to Word Processing	74
4.3	Text Editing and Formatting	77
4.4	Typical Processing Systems	81
4.4.1	Off-line Text-Editing Systems	81
4.4.2	Interactive Graphics-Editing Systems	92
4.5	Automatic Typesetting Systems	95
4.5.1	Typesetting Systems	95
◆ 4.5.2	Automatic Typefont Design	98

5	Text Compression	104
5.1	Statistical Language Characteristics	105
◆ 5.1.1	Frequency Considerations	105
◆ 5.1.2	Entropy Measurements	108
5.2	Rationale for Text Compression	111
5.3	Text-Compression Methods	114
◆ 5.3.1	Special-purpose Compression Systems	114
5.3.2	Basic Fixed-Length Codes	116
5.3.3	Restricted Variable-length Codes	119
5.3.4	Variable-length Codes	121
5.3.5	Word-Fragment Encoding	125

6	Text Encryption	131
6.1	Basic Cryptographic Concepts	131
6.2	Conventional Cryptographic Systems	135
6.3	Sample Cryptographic Ciphers	139
6.4	The Data Encryption Standard (DES)	146
◆ 6.5	Ciphers Based on Computationally Difficult Problems	149

- 7 File-Accessing Systems 159**
- 7.1 Basic Concepts 159
 - 7.2 Single-Key Searching: Sequential Search 161
 - 7.3 Single-Key Indexed Searches 162
 - 7.4 Tree Searching 167
 - 7.5 Balanced Search Trees 175
 - 7.6 Multiway Search Trees 183
 - 7.7 Hash-Table Access 192
 - 7.8 Indexed Searches for Multikey Access 201
 - 7.9 Bitmap Encoding for Multikey Access 206
 - ◆ 7.10 Multidimensional Access Structures 216

Part 3: Information-Retrieval Systems 227

- 8 Conventional Text-Retrieval Systems 229**
- 8.1 Database Management and Information Retrieval 229
 - 8.2 Text Retrieval Using Inverted Indexing Methods 231
 - 8.3 Extensions of the Inverted Index Operations 236
 - 8.3.1 Distance Constraints 236
 - 8.3.2 Term Weights 238
 - 8.3.3 Synonym Specifications 240
 - 8.3.4 Term Truncation 240
 - 8.4 Typical File Organization 243
 - ◆ 8.5 Optimization of Inverted-List Procedures 245
 - ◆ 8.5.1 Reducing the Number of Index Terms 245
 - ◆ 8.5.2 Quorum-level Searches 246
 - ◆ 8.5.3 Partial List Searching 248
 - 8.6 Text-scanning Systems 255
 - 8.6.1 General Considerations 255
 - 8.6.2 Elementary String Matching 256
 - 8.6.3 Fast String Matching 259
 - 8.7 Hardware Aids to Text Searching 266

- 9 Automatic Indexing 275**
- 9.1 Indexing Environment 275
 - 9.2 Indexing Aims 277
 - 9.3 Single-term Indexing Theories 279
 - 9.3.1 Term-frequency Considerations 279
 - 9.3.2 Term-discrimination Value 281
 - ◆ 9.3.3 Probabilistic Term Weighting 284
 - 9.4 Term Relationships in Indexing 290
 - 9.5 Term-phrase Formation 294
 - 9.6 Thesaurus-Group Generation 299
 - 9.7 A Blueprint for Automatic Indexing 303

10	Advanced Information-Retrieval Models	313
10.1	The Vector Space Model	312
◆	10.1.1 Basic Vector-processing Model	313
10.1.2	Vector Modifications	319
10.2	Automatic Document Classification	326
10.2.1	General Considerations	326
10.2.2	Hierarchical Cluster Generation	328
10.2.3	Heuristic Clustering Methods	338
10.2.4	Cluster Searching	341
◆	10.3 Probabilistic Retrieval Model	345
◆	10.4 Extended Boolean Retrieval Model	349
◆	10.4.1 Fuzzy Set Extensions	349
◆	10.4.2 Extended Boolean System	353
10.5	Integrated System for Processing Text and Data	362
10.6	Advanced Interface Systems	365

Part 4. Text Analysis and Language Processing 375

11	Language Analysis and Understanding	377
11.1	The Linguistic Approach	377
11.2	Dictionary Operations	379
11.2.1	Morphological Decomposition	379
11.2.2	Dictionary Types	382
11.3	Syntactic Analysis	386
11.3.1	Typical Syntactic-Analysis Systems	388
11.3.2	Semantic Grammars	398
11.4	Knowledge-based Processing	405
11.4.1	Knowledge Structures	405
11.4.2	Prospects for Knowledge-based Processing	408
11.5	Specialized Language Processing	410
◆	11.5.1 Robust Parsing	410
◆	11.5.2 Sublanguage Analysis	413
11.5.3	Natural-Language Interface to Information Systems	415

12	Automatic Text Transformations	425
12.1	Text Transformations	425
12.2	Automatic Writing Aids	426
12.2.1	Automatic Spelling Checkers	426
12.2.2	Automatic Spelling Correction	430
12.2.3	Syntax and Style Checking	436

- 12.3 Automatic Abstracting Systems 439
 - 12.3.1 Automatic Extracting 439
 - 12.3.2 Abstracting Based on Text Understanding 445
- ◆ 12.4 Automatic Text Generation 448
 - ◆ 12.4.1 Approaches to Text Generation 448
 - ◆ 12.4.2 Typical Text-generation Systems 451
- 12.5 Automatic Translation 456
 - 12.5.1 Main Approaches 456
 - 12.5.2 Typical Machine-translation Systems 461

13 Paperless Information Systems 471

- 13.1 Paperless Processing 471
- 13.2 Processing Complex Documents 473
- 13.3 Graphics Processing 477
 - 13.3.1 Basic Display Systems 477
 - 13.3.2 Object Transformations 479
 - 13.3.3 Picture Recognition 484
- 13.4 Speech Processing 486
 - 13.4.1 Speech Synthesis 487
 - 13.4.2 Speech Recognition 490
- 13.5 Automatic Teleconferencing Systems 495
- 13.6 Electronic Mail and Messages 497
- 13.7 Electronic Information Services 502
 - 13.7.1 Teletext 503
 - 13.7.2 Videotex 504
- 13.8 Electronic Publications and the Electronic Library 507

Author Index 517

Subject Index 523

Part I

The Information- Processing Environment

Chapter I

The Information Environment

II Automatic Information Processing

It has been claimed that we live in the information age, and our society is often called the information society. More information is produced and collected in our time than ever before: thousands of books, tens of thousands of journal articles, and innumerable informal studies and reports. Our capacity to absorb this information and use it in reaching intelligent decisions is stretched not only by the amount and variety of the available data, but also by the complex relationships among different types of information, and the resulting difficulties in interpreting the data.

Fortunately, although we are inundated by all sorts of information, improvements are being made in the ways in which information is stored and processed. In particular, modern information-processing equipment can organize and store large amounts of information and provide fast access to the stored records. Communications networks, used increasingly to reach the available information sources, also connect different information stores to large, often far-flung groups of users.

The use of modern computing equipment to process information has had a two-fold effect. On the one hand, it facilitates the generation, collection, and storage of more information, complicating the task of absorbing and using the available data. [1] On the other hand, modern equipment somewhat simplifies the problems of access to information by providing useful ways to search for and retrieve it.

This book deals with modern information processing, that is, the methods used to generate, analyze, store, retrieve, and handle information items using automatic equipment. Current capabilities in information processing are examined, and difficulties and conceptual problems in analyzing and understanding information are described. By distinguishing relatively routine tasks from more experimental, laboratory-type endeavors, and by considering future developments, the book also outlines the information-processing world of the future.

1.2 Types of Information

Information can take three forms: written texts, spoken utterances, and graphs and images. Text, the basic medium for formal communications between human beings, consists of notes, messages, letters, memoranda, books, newspapers, magazines, and so on. Speech is more informal than text and, unlike text, is also accessible to people who cannot read or write. Graphs and images may accompany written texts, but can also be used alone as illustrations, displays, movies, or paintings.

In dealing with these information types it is useful to consider two principal aspects of information processing. The first area is the *technical* problem of information representation and manipulation, including methods of introducing and storing information in computers, and of transferring the data and making them accessible to interested users. The second area relates to the *semantic* and *behavioral* aspects of information processing: the accuracy with which the stored information conveys intended meanings, and the effectiveness with which it affects users' conduct as intended.

From a technical point of view, stored information can be treated simply as collections of disconnected elements — for example, individual words in given texts, individual characters in particular words, or picture elements in graphs and pictures. For processing purposes, the information elements are not assumed to convey specific meanings or to be tied to particular contexts. Thus a text can be reproduced or copied without the text content ever being considered. In actual fact, however, the information elements do carry meaning, and are expected to generate specific responses by the information users. Ultimately the meaning of the information tends to be more important than the form of representation and the manner in which the data are manipulated.