

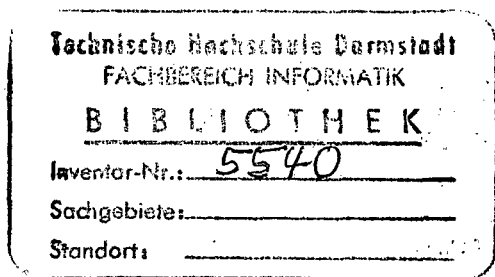
# Introduction to Modern Information Retrieval

Gerard Salton

*Professor of Computer Science  
Cornell University*

Michael J. McGill

*Associate Professor of Information Studies  
Syracuse University*



INTERNATIONAL STUDENT EDITION

McGRAW-HILL INTERNATIONAL BOOK COMPANY

Auckland Bogotá Guatemala Hamburg Johannesburg Lisbon  
London Madrid Mexico New Delhi Panama Paris  
San Juan São Paulo Singapore Sydney Tokyo

# Contents

Preface	xi
CHAPTER 1 Information Retrieval: An Introduction	1
0 Preview	1
1 Overview	1
2 Changing Technology	5
3 Information System Types	7
A Information Retrieval Systems	7
B Data Base Management Systems	8
C Management Information Systems	8
D Decision Support Systems	9
E Question-Answering Systems	9
4 Functional Approach to Information Retrieval	10
5 Simple File Structures	12
A Linear Lists	12
B Ordered Sequential Files	13
*C Indexed Files	16
6 Summary	21
CHAPTER 2 Systems Based on Inverted Files	24
0 Preview	24
1 General Considerations	24

A Boolean Expressions	25
B Order of Operations	26
2 Adjacency and Term Frequency Features	28
A Adjacency Operations	28
B Frequency Information	30
3 Commercial Inverted File Systems	30
A The DIALOG System	30
*B The STAIRS System	34
C The Bibliographic Retrieval Services (BRS) System	41
D The MEDLARS System	42
E The ORBIT System	45
F The Information Bank	45
G The LEXIS System	46
4 Enhancements of Basic Retrieval Strategy	46
 CHAPTER 3 Text Analysis and Automatic Indexing	 52
0 Preview	52
1 Indexing Environment	53
2 Manual and Automatic Indexing	55
3 Automatic Term Extraction and Weighting	59
A General Considerations	59
*B The Inverse Document Frequency Weight	63
**C The Signal-Noise Ratio	63
*D The Term Discrimination Value	66
4 A Simple Automatic Indexing Process	71
5 Automatic Term Association and Use of Context	75
A Thesaurus Rules	75
*B Automatic Thesaurus Construction	78
C Thesaurus Use	81
D Construction of Term Phrases	84
E Automatic Sentence Extraction	87
6 Some Theoretical Approaches	90
*A The Use of Linguistic Methods	90
*B Fragment Encoding	93
**C Probabilistic Information Retrieval	94
7 Automatic Indexing Experiments	99
 CHAPTER 4 The SMART and SIRE Experimental Retrieval Systems	 118
0 Preview	118
1 Introduction	118
2 The SMART System Environment	120
*A Vector Representation and Similarity Computation	120
*B Vector Manipulation	123
C Vector Generation	127
3 SMART System Procedures	130
*A Automatic Indexing	130
*B Automatic Document Classification	137
*C Relevance Feedback Operations	140
*D Dynamic Document Space	145

4 Automatic Enhancements of Conventional Retrieval	146
*A Document Ranking and Term Weighting	146
*B Retrieval through Man-Machine Dialogue and Local Clustering	151
<b>CHAPTER 5 Retrieval Evaluation</b>	<b>157</b>
0 Preview	157
1 Introduction	158
2 Evaluation of Retrieval Effectiveness	159
A System Components	159
B Evaluation Viewpoints and the Relevance Problem	161
*C The Computation of Recall and Precision	164
3 Measures of Retrieval Effectiveness	172
A Measurement Problems	172
*B Recall, Precision, and Fallout	174
**C Single-Valued Measures	177
**D Utility Measure	184
4 Evaluation of System Cost and Efficiency	186
A System Tradeoffs	186
**B Cost Analysis	187
5 Summary	191
<b>CHAPTER 6 Retrieval Refinements</b>	<b>199</b>
0 Preview	199
1 Introduction	200
*2 Vector Similarity Functions	201
3 Term Weighting Systems	204
A Principal Weighting Strategies	204
*B Evaluation of Weighting Systems	207
**C Term Weighting in Boolean Query Systems	211
4 File Clustering	215
*A Main Considerations	215
*B Classification Methods	220
*C Cluster Search Evaluation	222
**D Automatic Pseudoclassification	227
5 Dynamic Query Adjustment	236
A General Considerations	236
*B Feedback Theory	238
*C Feedback Variations	240
D Dynamic Document Space	244
6 Citation Processing	246
A Basic Citation Properties	246
*B Main Citation Usage	247
7 Summary	250
<b>CHAPTER 7 Natural Language Processing</b>	<b>257</b>
0 Preview	257
1 Components of Natural Language Systems	258
A Interest in Natural Language Processing	258

B Levels of Language Processing	259
C Language Understanding Systems	261
2 Language Processing and Information Retrieval	266
3 Syntactic Analysis Systems	267
*A Phrase Structure Grammars	268
*B Transformational Grammars	273
**C Augmented Transition Network Grammars	276
4 Syntactic Analysis in Information Retrieval	284
5 Linguistic Methods in Question Answering	287
**A Knowledge Representation	287
B Question-Answering Environment	291
*C Linguistic Features in Question Answering	292
6 Summary	297
 CHAPTER 8 Access to Information: Hardware and Software Approaches	 303
0 Preview	303
1 Conventional Storage Devices	303
A Punched Cards	304
B Magnetic Tape	306
C Magnetic Disks	307
D Random Access Storage Devices	308
E Data Cell	309
F Access to Storage	310
2 Hardware Enhancement of Retrieval	312
A Microprocessors and Processing Chips	312
B General Characteristics of Retrieval Hardware	314
C Parallel Processors	316
*D Associative Processors	317
*E Fast Computations Using Array Processors	320
*F Content Addressable Segment Sequential Memory (CASSM)	322
*G Relational Associative Processors (RAP)	324
*H Data Base Computer (DBC)	326
I Other Special-Purpose Devices	328
3 Text Access Methods	329
*A Dictionary Search Methods for Static Files	329
*B Dictionary Search Methods for Dynamic Files	333
*C Multiple Key Dictionary Search	338
D Text Scanning Machines	339
**E String Matching Using the Finite State Automaton Model	340
**F The Boyer and Moore String Matching Method	345
4 Summary	348
 CHAPTER 9 Data Management Systems	 354
0 Preview	354
1 Types of Information Systems	355
A Information Retrieval and Question Answering	355
B Data Management Systems	357
2 The Structure of Data Base Management Systems	359
A Basic Concepts	359
B Structure of Information Items	362

*C The Relational Data Base Model	365
*D The Hierarchical Data Base Model	370
*E The Network Data Base Model	377
3 Query Processing	380
*A Query Language Types	380
*B Processing Strategies	386
4 Data Quality	390
*A Integrity and Security	390
**B Concurrent Data Base Operations	394
**C Restart and Recovery Methods	398
**D Distributed Data Bases	399
5 Summary	401

CHAPTER 10 Future Directions in Information Retrieval	408
0 Preview	408
1 Introduction	409
2 Technological Developments	410
A Automatic Document Input	410
B Optical Storage	413
3 Information Theories and Models	418
A Natural Language Processing	418
**B Fuzzy Set Theory	421
**C Term Dependency Models	422
*D Composite Document Representations	425
4 Advanced Information Systems	426
A Mixed Information Retrieval Systems	426
B Personal Computing and Paperless Information Systems	428
5 Conclusion	430

Indexes	437
Name Index	
Subject Index	