

Contents

Preface	v
Acknowledgements	vii
Biographies	xvii
1 Introduction	1
1.1 Motivation	1
1.1.1 Information versus Data Retrieval	1
1.1.2 Information Retrieval at the Center of the Stage	2
1.1.3 Focus of the Book	3
1.2 Basic Concepts	3
1.2.1 The User Task	4
1.2.2 Logical View of the Documents	5
1.3 Past, Present, and Future	6
1.3.1 Early Developments	6
1.3.2 Information Retrieval in the Library	7
1.3.3 The Web and Digital Libraries	7
1.3.4 Practical Issues	8
1.4 The Retrieval Process	9
1.5 Organization of the Book	10
1.5.1 Book Topics	11
1.5.2 Book Chapters	12
1.6 How to Use this Book	15
1.6.1 Teaching Suggestions	15
1.6.2 The Book's Web Page	16
1.7 Bibliographic Discussion	17
2 Modeling	19
2.1 Introduction	19
2.2 A Taxonomy of Information Retrieval Models	20
2.3 Retrieval: Ad hoc and Filtering	21

2.4	A Formal Characterization of IR Models	23
2.5	Classic Information Retrieval	24
2.5.1	Basic Concepts	24
2.5.2	Boolean Model	25
2.5.3	Vector Model	27
2.5.4	Probabilistic Model	30
2.5.5	Brief Comparison of Classic Models	34
2.6	Alternative Set Theoretic Models	34
2.6.1	Fuzzy Set Model	34
2.6.2	Extended Boolean Model	38
2.7	Alternative Algebraic Models	41
2.7.1	Generalized Vector Space Model	41
2.7.2	Latent Semantic Indexing Model	44
2.7.3	Neural Network Model	46
2.8	Alternative Probabilistic Models	48
2.8.1	Bayesian Networks	48
2.8.2	Inference Network Model	49
2.8.3	Belief Network Model	56
2.8.4	Comparison of Bayesian Network Models	59
2.8.5	Computational Costs of Bayesian Networks	60
2.8.6	The Impact of Bayesian Network Models	61
2.9	Structured Text Retrieval Models	61
2.9.1	Model Based on Non-Overlapping Lists	62
2.9.2	Model Based on Proximal Nodes	63
2.10	Models for Browsing	65
2.10.1	Flat Browsing	65
2.10.2	Structure Guided Browsing	66
2.10.3	The Hypertext Model	66
2.11	Trends and Research Issues	69
2.12	Bibliographic Discussion	69
3	Retrieval Evaluation	73
3.1	Introduction	73
3.2	Retrieval Performance Evaluation	74
3.2.1	Recall and Precision	75
3.2.2	Alternative Measures	82
3.3	Reference Collections	84
3.3.1	The TREC Collection	84
3.3.2	The CACM and ISI Collections	91
3.3.3	The Cystic Fibrosis Collection	94
3.4	Trends and Research Issues	96
3.5	Bibliographic Discussion	96
4	Query Languages	99
4.1	Introduction	99
4.2	Keyword-Based Querying	100

4.2.1	Single-Word Queries	100
4.2.2	Context Queries	101
4.2.3	Boolean Queries	102
4.2.4	Natural Language	103
4.3	Pattern Matching	104
4.4	Structural Queries	106
4.4.1	Fixed Structure	108
4.4.2	Hypertext	108
4.4.3	Hierarchical Structure	109
4.5	Query Protocols	113
4.6	Trends and Research Issues	114
4.7	Bibliographic Discussion	116
5	Query Operations	117
5.1	Introduction	117
5.2	User Relevance Feedback	118
5.2.1	Query Expansion and Term Reweighting for the Vector Model	118
5.2.2	Term Reweighting for the Probabilistic Model	120
5.2.3	A Variant of Probabilistic Term Reweighting	121
5.2.4	Evaluation of Relevance Feedback Strategies	122
5.3	Automatic Local Analysis	123
5.3.1	Query Expansion Through Local Clustering	124
5.3.2	Query Expansion Through Local Context Analysis	129
5.4	Automatic Global Analysis	131
5.4.1	Query Expansion based on a Similarity Thesaurus	131
5.4.2	Query Expansion based on a Statistical Thesaurus	134
5.5	Trends and Research Issues	137
5.6	Bibliographic Discussion	138
6	Text and Multimedia Languages and Properties	141
6.1	Introduction	141
6.2	Metadata	142
6.3	Text	144
6.3.1	Formats	144
6.3.2	Information Theory	145
6.3.3	Modeling Natural Language	145
6.3.4	Similarity Models	148
6.4	Markup Languages	149
6.4.1	SGML	149
6.4.2	HTML	152
6.4.3	XML	154
6.5	Multimedia	156
6.5.1	Formats	157
6.5.2	Textual Images	158
6.5.3	Graphics and Virtual Reality	159

6.5.4	HyTime	159
6.6	Trends and Research Issues	160
6.7	Bibliographic Discussion	162
7	Text Operations	163
7.1	Introduction	163
7.2	Document Preprocessing	165
7.2.1	Lexical Analysis of the Text	165
7.2.2	Elimination of Stopwords	167
7.2.3	Stemming	168
7.2.4	Index Terms Selection	169
7.2.5	Thesauri	170
7.3	Document Clustering	173
7.4	Text Compression	173
7.4.1	Motivation	173
7.4.2	Basic Concepts	175
7.4.3	Statistical Methods	176
7.4.4	Dictionary Methods	183
7.4.5	Inverted File Compression	184
7.5	Comparing Text Compression Techniques	186
7.6	Trends and Research Issues	188
7.7	Bibliographic Discussion	189
8	Indexing and Searching	191
8.1	Introduction	191
8.2	Inverted Files	192
8.2.1	Searching	195
8.2.2	Construction	196
8.3	Other Indices for Text	199
8.3.1	Suffix Trees and Suffix Arrays	199
8.3.2	Signature Files	205
8.4	Boolean Queries	207
8.5	Sequential Searching	209
8.5.1	Brute Force	209
8.5.2	Knuth-Morris-Pratt	210
8.5.3	Boyer-Moore Family	211
8.5.4	Shift-Or	212
8.5.5	Suffix Automaton	213
8.5.6	Practical Comparison	214
8.5.7	Phrases and Proximity	215
8.6	Pattern Matching	215
8.6.1	String Matching Allowing Errors	216
8.6.2	Regular Expressions and Extended Patterns	219
8.6.3	Pattern Matching Using Indices	220
8.7	Structural Queries	222
8.8	Compression	222

8.8.1	Sequential Searching	223
8.8.2	Compressed Indices	224
8.9	Trends and Research Issues	226
8.10	Bibliographic Discussion	227
9	Parallel and Distributed IR	229
9.1	Introduction	229
9.1.1	Parallel Computing	230
9.1.2	Performance Measures	231
9.2	Parallel IR	232
9.2.1	Introduction	232
9.2.2	MIMD Architectures	233
9.2.3	SIMD Architectures	240
9.3	Distributed IR	249
9.3.1	Introduction	249
9.3.2	Collection Partitioning	251
9.3.3	Source Selection	252
9.3.4	Query Processing	253
9.3.5	Web Issues	254
9.4	Trends and Research Issues	255
9.5	Bibliographic Discussion	256
10	User Interfaces and Visualization	257
10.1	Introduction	257
10.2	Human-Computer Interaction	258
10.2.1	Design Principles	258
10.2.2	The Role of Visualization	259
10.2.3	Evaluating Interactive Systems	261
10.3	The Information Access Process	262
10.3.1	Models of Interaction	262
10.3.2	Non-Search Parts of the Information Access Process	265
10.3.3	Earlier Interface Studies	266
10.4	Starting Points	267
10.4.1	Lists of Collections	267
10.4.2	Overviews	268
10.4.3	Examples, Dialogs, and Wizards	276
10.4.4	Automated Source Selection	278
10.5	Query Specification	278
10.5.1	Boolean Queries	279
10.5.2	From Command Lines to Forms and Menus	280
10.5.3	Faceted Queries	281
10.5.4	Graphical Approaches to Query Specification	282
10.5.5	Phrases and Proximity	286
10.5.6	Natural Language and Free Text Queries	287
10.6	Context	289
10.6.1	Document Surrogates	289

10.6.2	Query Term Hits Within Document Content	289
10.6.3	Query Term Hits Between Documents	293
10.6.4	SuperBook: Context via Table of Contents	296
10.6.5	Categories for Results Set Context	297
10.6.6	Using Hyperlinks to Organize Retrieval Results	299
10.6.7	Tables	301
10.7	Using Relevance Judgements	303
10.7.1	Interfaces for Standard Relevance Feedback	304
10.7.2	Studies of User Interaction with Relevance Feedback Systems	305
10.7.3	Fetching Relevant Information in the Background	307
10.7.4	Group Relevance Judgements	308
10.7.5	Pseudo-Relevance Feedback	308
10.8	Interface Support for the Search Process	309
10.8.1	Interfaces for String Matching	309
10.8.2	Window Management	311
10.8.3	Example Systems	312
10.8.4	Examples of Poor Use of Overlapping Windows	317
10.8.5	Retaining Search History	317
10.8.6	Integrating Scanning, Selection, and Querying	318
10.9	Trends and Research Issues	321
10.10	Bibliographic Discussion	322
11	Multimedia IR: Models and Languages	325
11.1	Introduction	325
11.2	Data Modeling	328
11.2.1	Multimedia Data Support in Commercial DBMSs	329
11.2.2	The MULTOS Data Model	331
11.3	Query Languages	334
11.3.1	Request Specification	335
11.3.2	Conditions on Multimedia Data	335
11.3.3	Uncertainty, Proximity, and Weights in Query Expressions	337
11.3.4	Some Proposals	338
11.4	Trends and Research Issues	341
11.5	Bibliographic Discussion	342
12	Multimedia IR: Indexing and Searching	345
12.1	Introduction	345
12.2	Background — Spatial Access Methods	347
12.3	A Generic Multimedia Indexing Approach	348
12.4	One-dimensional Time Series	353
12.4.1	Distance Function	353
12.4.2	Feature Extraction and Lower-bounding	353
12.4.3	Experiments	355
12.5	Two-dimensional Color Images	357

12.5.1	Image Features and Distance Functions	357
12.5.2	Lower-bounding	358
12.5.3	Experiments	360
12.6	Automatic Feature Extraction	360
12.7	Trends and Research Issues	361
12.8	Bibliographic Discussion	363
13	Searching the Web	367
13.1	Introduction	367
13.2	Challenges	368
13.3	Characterizing the Web	369
13.3.1	Measuring the Web	369
13.3.2	Modeling the Web	371
13.4	Search Engines	373
13.4.1	Centralized Architecture	373
13.4.2	Distributed Architecture	375
13.4.3	User Interfaces	377
13.4.4	Ranking	380
13.4.5	Crawling the Web	382
13.4.6	Indices	383
13.5	Browsing	384
13.5.1	Web Directories	384
13.5.2	Combining Searching with Browsing	386
13.5.3	Helpful Tools	387
13.6	Metasearchers	387
13.7	Finding the Needle in the Haystack	389
13.7.1	User Problems	389
13.7.2	Some Examples	390
13.7.3	Teaching the User	391
13.8	Searching using Hyperlinks	392
13.8.1	Web Query Languages	392
13.8.2	Dynamic Search and Software Agents	393
13.9	Trends and Research Issues	393
13.10	Bibliographic Discussion	395
14	Libraries and Bibliographical Systems	397
14.1	Introduction	397
14.2	Online IR Systems and Document Databases	398
14.2.1	Databases	399
14.2.2	Online Retrieval Systems	403
14.2.3	IR in Online Retrieval Systems	404
14.2.4	'Natural Language' Searching	406
14.3	Online Public Access Catalogs (OPACs)	407
14.3.1	OPACs and Their Content	408
14.3.2	OPACs and End Users	410
14.3.3	OPACs: Vendors and Products	410

14.3.4	Alternatives to Vendor OPACs	410
14.4	Libraries and Digital Library Projects	412
14.5	Trends and Research Issues	412
14.6	Bibliographic Discussion	413
15	Digital Libraries	415
15.1	Introduction	415
15.2	Definitions	417
15.3	Architectural Issues	418
15.4	Document Models, Representations, and Access	420
15.4.1	Multilingual Documents	420
15.4.2	Multimedia Documents	421
15.4.3	Structured Documents	421
15.4.4	Distributed Collections	422
15.4.5	Federated Search	424
15.4.6	Access	424
15.5	Prototypes, Projects, and Interfaces	425
15.5.1	International Range of Efforts	427
15.5.2	Usability	428
15.6	Standards	429
15.6.1	Protocols and Federation	429
15.6.2	Metadata	430
15.7	Trends and Research Issues	431
15.8	Bibliographical Discussion	432
	Appendix: Porter's Algorithm	433
	Glossary	437
	References	455
	Index	501