

Contents

Preface to the Second Edition	xix
Preface to the First Edition	xxi
Authors' Acknowledgements to the Second Edition	xxiii
Authors' Acknowledgements to the First Edition	xxv
Publishers' Acknowledgements	xxvii
1 Introduction	1
1.1 Information Retrieval	1
1.1.1 Early Developments	1
1.1.2 Information Retrieval in Libraries and Digital Libraries	3
1.1.3 IR at the Center of the Stage	3
1.2 The IR Problem	3
1.2.1 The User's Task	4
1.2.2 Information versus Data Retrieval	5
1.3 The IR System	5
1.3.1 Software Architecture of the IR System	5
1.3.2 The Retrieval and Ranking Processes	7
1.4 The Web	8
1.4.1 A Brief History	8
1.4.2 The e-Publishing Era	9
1.4.3 How the Web Changed Search	10
1.4.4 Practical Issues on the Web	12
1.5 Organization of the Book	12
1.5.1 Focus of the Book	12
1.5.2 Book Contents	13
1.6 The Book Web Site: A Teaching Resource	16
1.7 Bibliographic Discussion	17
2 User Interfaces for Search	21
<i>by Marti Hearst</i>	
2.1 Introduction	21
2.2 How People Search	21

2.2.1	Information Lookup versus Exploratory Search	22
2.2.2	Classic versus Dynamic Model of Information Seeking	23
2.2.3	Navigation versus Search	24
2.2.4	Observations of the Search Process	24
2.3	Search Interfaces Today	25
2.3.1	Getting Started	25
2.3.2	Query Specification	26
2.3.3	Query Specification Interfaces	27
2.3.4	Retrieval Results Display	29
2.3.5	Query Reformulation	32
2.3.6	Organizing Search Results	35
2.4	Visualization in Search Interfaces	40
2.4.1	Visualizing Boolean Syntax	42
2.4.2	Visualizing Query Terms within Retrieval Results	43
2.4.3	Visualizing Relationships Among Words and Documents	47
2.4.4	Visualization for Text Mining	49
2.5	Design and Evaluation of Search Interfaces	50
2.6	Trends and Research Issues	54
2.7	Bibliographic Discussion	54
3	Modeling	57
3.1	IR Models	57
3.1.1	Modeling and Ranking	57
3.1.2	Characterization of an IR Model	58
3.1.3	A Taxonomy of IR Models	59
3.2	Classic Information Retrieval	61
3.2.1	Basic Concepts	61
3.2.2	The Boolean Model	64
3.2.3	Term Weighting	66
3.2.4	TF-IDF Weights	68
3.2.5	Document Length Normalization	75
3.2.6	The Vector Model	77
3.2.7	The Probabilistic Model	79
3.2.8	Brief Comparison of Classic Models	86
3.3	Alternative Set Theoretic Models	87
3.3.1	Set-Based Model	87
3.3.2	Extended Boolean Model	92
3.3.3	Fuzzy Set Model	95
3.4	Alternative Algebraic Models	98
3.4.1	Generalized Vector Space Model	98
3.4.2	Latent Semantic Indexing Model	101
3.4.3	Neural Network Model	102
3.5	Alternative Probabilistic Models	104
3.5.1	BM25	104
3.5.2	Language Models	107
3.5.3	Divergence from Randomness	113
3.5.4	Bayesian Network Models	116
3.6	Other Models	124

3.6.1	The Hypertext Model	124
3.6.2	Web based Models	125
3.6.3	Structured Text Retrieval	126
3.6.4	Multimedia Retrieval	126
3.6.5	Enterprise and Vertical Search	126
3.7	Trends and Research Issues	127
3.8	Bibliographic Discussion	128
4	Retrieval Evaluation	131
4.1	Introduction	131
4.2	The Cranfield Paradigm	132
4.2.1	A Brief History	132
4.2.2	Reference Collections	134
4.3	Retrieval Metrics	134
4.3.1	Precision and Recall	135
4.3.2	Single Value Summaries: P@n, MAP, MRR, F	139
4.3.3	User-Oriented Measures	144
4.3.4	DCG: Discounted Cumulated Gain	145
4.3.5	BPREF: Binary Preferences	150
4.3.6	Rank Correlation Metrics	153
4.4	Reference Collections	158
4.4.1	The TREC Collections	159
4.4.2	Other Reference Collections	166
4.4.3	Other Small Test Collections	167
4.5	User-Based Evaluation	168
4.5.1	Human Experimentation in the Lab	168
4.5.2	Side-by-Side Panels	168
4.5.3	A/B Testing	169
4.5.4	Crowdsourcing	170
4.5.5	Evaluation using Clickthrough Data	171
4.6	Practical Caveats	173
4.7	Trends and Research Issues	174
4.8	Bibliographic Discussion	174
5	Relevance Feedback and Query Expansion	177
5.1	Introduction	177
5.2	A Framework for Feedback Methods	178
5.3	Explicit Relevance Feedback	180
5.3.1	Relevance Feedback for the Vector Model: Rocchio Method	181
5.3.2	Relevance Feedback for the Probabilistic Model	183
5.3.3	Evaluation of Relevance Feedback	184
5.4	Explicit Feedback Through Clicks	185
5.4.1	Eye Tracking and Relevance Judgements	185
5.4.2	User Behavior	186
5.4.3	Clicks as a Metric of User Preferences	187
5.5	Implicit Feedback Through Local Analysis	190
5.5.1	Implicit Feedback Through Local Clustering	190
5.5.2	Implicit Feedback through Local Context Analysis	193

5.6	Implicit Feedback Through Global Analysis	195
5.6.1	Query Expansion based on a Similarity Thesaurus	195
5.6.2	Query Expansion based on a Statistical Thesaurus	198
5.7	Trends and Research Issues	200
5.8	Bibliographic Discussion	200
6	Documents: Languages & Properties	203
	<i>with Gonzalo Navarro and Nivio Ziviani</i>	
6.1	Introduction	203
6.2	Metadata	205
6.3	Document Formats	206
6.3.1	Text	206
6.3.2	Multimedia	207
6.3.3	Graphics and Virtual Reality	208
6.4	Markup Languages	208
6.4.1	SGML	209
6.4.2	HTML	211
6.4.3	XML	214
6.4.4	RDF: Resource Description Framework	216
6.4.5	HyTime	217
6.5	Text Properties	218
6.5.1	Information Theory	218
6.5.2	Modeling Natural Language	219
6.5.3	Text Similarity	222
6.6	Document Preprocessing	223
6.6.1	Lexical Analysis of the Text	224
6.6.2	Elimination of Stopwords	226
6.6.3	Stemming	226
6.6.4	Keyword Selection	227
6.6.5	Thesauri	228
6.7	Organizing Documents	231
6.7.1	Taxonomies	231
6.7.2	Folksonomies	232
6.8	Text Compression	233
6.8.1	Basic Concepts	234
6.8.2	Statistical Methods	234
6.8.3	Statistical Methods: Modeling	235
6.8.4	Statistical Methods: Coding	238
6.8.5	Dictionary Methods	245
6.8.6	Preprocessing for Compression	246
6.8.7	Comparing Text Compression Techniques	248
6.8.8	Structured Text Compression	249
6.9	Trends and Research Issues	250
6.10	Bibliographical Discussion	253
7	Queries: Languages & Properties	255
	<i>with Gonzalo Navarro</i>	
7.1	Query Languages	255

7.1.1	Keyword-based Querying	256
7.1.2	Beyond Keywords	259
7.1.3	Structural Queries	262
7.1.4	Query Protocols	265
7.2	Query Properties	267
7.2.1	Characterizing Web Queries	267
7.2.2	User Search Behavior	269
7.2.3	Query Intent	270
7.2.4	Query Topic	272
7.2.5	Query Sessions and Missions	273
7.2.6	Query Difficulty	274
7.3	Trends and Research Issues	278
7.4	Bibliographical Discussion	279
8	Text Classification	281
	<i>with Marcos Gonçalves</i>	
8.1	Introduction	281
8.2	A Characterization of Text Classification	282
8.2.1	Machine Learning	282
8.2.2	The Text Classification Problem	283
8.2.3	Text Classification Algorithms	284
8.3	Unsupervised Algorithms	286
8.3.1	Clustering	286
8.3.2	Naive Text Classification	290
8.4	Supervised Algorithms	291
8.4.1	Decision Trees	294
8.4.2	The k-NN Classifier	299
8.4.3	The Rocchio Classifier	300
8.4.4	Probabilistic Naive Bayes Document Classification	303
8.4.5	The SVM Classifier	306
8.4.6	Ensemble Classifiers	316
8.4.7	Final Remarks on Supervised Algorithms	319
8.5	Feature Selection or Dimensionality Reduction	320
8.5.1	Term-Class Incidence Table	321
8.5.2	Term Document Frequency	322
8.5.3	TF-IDF Weights	322
8.5.4	Mutual Information	323
8.5.5	Information Gain	323
8.5.6	Chi Square	324
8.5.7	Impact of Feature Selection	325
8.6	Evaluation Metrics	325
8.6.1	Contingency Table	325
8.6.2	Accuracy and Error	326
8.6.3	Precision and Recall	327
8.6.4	F-measure and F_1	327
8.6.5	Cross-Validation	329
8.6.6	Standard Collections	329
8.7	Organizing the Classes – Building Taxonomies	330

8.8	Trends and Research Issues	333
8.9	Bibliographic Discussion	334
9	Indexing and Searching	337
	<i>with Gonzalo Navarro</i>	
9.1	Introduction	337
9.2	Inverted Indexes	340
9.2.1	Basic Concepts	340
9.2.2	Full Inverted Indexes	341
9.2.3	Searching	345
9.2.4	Ranking	348
9.2.5	Construction	351
9.2.6	Compressed Inverted Indexes	354
9.2.7	Structural Queries	357
9.3	Signature Files	357
9.4	Suffix Trees and Suffix Arrays	360
9.4.1	Structure: Tries and Suffix Trees	361
9.4.2	Searching for Simple Strings	362
9.4.3	Searching for Complex Patterns	363
9.4.4	Construction	365
9.4.5	Compressed Suffix Arrays	367
9.5	Sequential Searching	372
9.5.1	Simple Strings: Horspool	373
9.5.2	Complex Patterns: Automata and Bit-Parallelism	375
9.5.3	Faster Bit-Parallel Algorithms	379
9.5.4	Regular Expressions	382
9.5.5	Multiple Patterns	384
9.5.6	Approximate Searching	385
9.5.7	Searching Compressed Text	389
9.6	Multi-dimensional Indexing	391
9.7	Trends and Research Issues	393
9.8	Bibliographic Discussion	394
10	Parallel and Distributed IR	399
	<i>with Eric Brown</i>	
10.1	Introduction	399
10.2	A Taxonomy of Distributed IR Systems	402
10.3	Data Partitioning	404
10.3.1	Collection Partitioning	405
10.3.2	Collection Selection	407
10.3.3	Inverted Index Partitioning	409
10.3.4	Partitioning other Indexes	413
10.4	Parallel IR	414
10.4.1	Introduction	414
10.4.2	Parallel IR on MIMD Architectures	416
10.4.3	Parallel IR on SIMD Architectures	418
10.5	Cluster-based IR	423
10.6	Distributed IR	424

10.6.1	Introduction	424
10.6.2	Indexing	428
10.6.3	Query Processing	431
10.6.4	Web Issues	437
10.7	Federated Search	438
10.8	Retrieval in Peer-to-Peer Networks	440
10.9	Trends and Research Issues	444
10.10	Bibliographic Discussion	445
11	Web Retrieval	447
	<i>with Yoelle Maarek</i>	
11.1	Introduction	447
11.2	A Challenging Problem	449
11.3	The Web	451
11.3.1	Characteristics	451
11.3.2	Structure of the Web Graph	452
11.3.3	Modeling the Web	454
11.3.4	Link Analysis	456
11.4	Search Engine Architectures	458
11.4.1	Basic Architecture	458
11.4.2	Cluster-based Architecture	459
11.4.3	Caching	462
11.4.4	Multiple Indexes	464
11.4.5	Distributed Architectures	466
11.5	Search Engine Ranking	468
11.5.1	Ranking Signals	469
11.5.2	Link-based Ranking	470
11.5.3	Simple Ranking Functions	473
11.5.4	Learning to Rank	473
11.5.5	Learning the Ranking Function	474
11.5.6	Quality Evaluation	475
11.5.7	Web Spam	476
11.6	Managing Web Data	477
11.6.1	Assigning Identifiers to Documents	477
11.6.2	Metadata	478
11.6.3	Compressing the Web Graph	478
11.6.4	Handling Duplicated Data	479
11.7	Search Engine User Interaction	480
11.7.1	The Search Rectangle Paradigm	481
11.7.2	The Search Engine Result Page	488
11.7.3	Educating the User	497
11.8	Browsing	498
11.8.1	Flat Browsing	499
11.8.2	Structure Guided Browsing and Web Directories	499
11.9	Beyond Browsing	501
11.9.1	Hypertext and the Web	501
11.9.2	Combining Searching with Browsing	501
11.9.3	Web Query Languages	503

11.9.4	Dynamic Search	503
11.10	Related Problems	504
11.10.1	Computational Advertising	504
11.10.2	Web Mining	506
11.10.3	Metasearch	508
11.11	Trends and Research Issues	509
11.11.1	Beyond Static Text Data	509
11.11.2	Current Challenges	511
11.12	Bibliographical Discussion	513
12	Web Crawling	515
	<i>with Carlos Castillo</i>	
12.1	Introduction	515
12.2	Applications of a Web Crawler	517
12.2.1	General Web Search	517
12.2.2	Topical Crawling	518
12.2.3	Web Characterization	518
12.2.4	Mirroring	518
12.2.5	Web Site Analysis	519
12.3	A Taxonomy of Crawlers	519
12.3.1	Types of Web Pages	520
12.4	Architecture and Implementation	521
12.4.1	Crawler Architecture	521
12.4.2	Practical Issues	523
12.4.3	Parallel Crawling	526
12.5	Scheduling Algorithms	527
12.5.1	Selection Policy	528
12.5.2	Revisit Policy	530
12.5.3	Politeness Policy	535
12.5.4	Combining Policies	538
12.6	Evaluation	539
12.6.1	Evaluating Network Usage	539
12.6.2	Evaluating Long-term Scheduling	540
12.7	Trends and Research Issues	541
12.7.1	Crawling the "Hidden" Web	541
12.7.2	Crawling with the Help of Web Sites	542
12.7.3	Distributed Crawling	543
12.8	Bibliographic Discussion	543
13	Structured Text Retrieval	545
	<i>with Mounia Lalmas</i>	
13.1	Introduction	545
13.2	Structuring Power	546
13.2.1	Explicit vs. Implicit Structure	546
13.2.2	Static vs. Dynamic Structure	547
13.2.3	Single Hierarchy vs. Multiple Hierarchies	548
13.3	Early Text Retrieval Models	549
13.3.1	Model Based on Non-Overlapping Lists	549

13.3.2	Model Based on Proximal Nodes	550
13.3.3	Ranking Structured Text Results	551
13.4	XML Retrieval	551
13.4.1	Challenges in XML Retrieval	551
13.4.2	Indexing Strategies	553
13.4.3	Ranking Strategies	554
13.4.4	Removing Overlaps	565
13.5	XML Retrieval Evaluation	566
13.5.1	Document Collections	566
13.5.2	Topics	567
13.5.3	Retrieval Tasks	568
13.5.4	Relevance	569
13.5.5	Measures	571
13.6	Query Languages	573
13.6.1	Characteristics	574
13.6.2	Classification of XML Query Languages	575
13.6.3	Examples of XML Query Languages	577
13.7	Trends and Research Issues	582
13.8	Bibliographic Discussion	585
14	Multimedia Information Retrieval	587
	<i>by Dulce Poncelaón and Malcolm Slaney</i>	
14.1	Introduction	587
14.1.1	What is Multimedia?	587
14.1.2	Multimedia IR	588
14.1.3	Text IR versus Multimedia IR	589
14.2	The Challenges	589
14.2.1	The Semantic Gap	589
14.2.2	Feature Ambiguity	591
14.2.3	Machine-generated Data	591
14.3	Content-based Image Retrieval	592
14.3.1	Color-Based Retrieval	593
14.3.2	Texture	593
14.3.3	Salient Points	596
14.4	Audio and Music Retrieval	597
14.4.1	Fingerprinting	598
14.4.2	Speech Recognition	599
14.4.3	Speaker Identification	601
14.4.4	Spoken Document Retrieval	602
14.4.5	Audio Basics	602
14.5	Retrieving and Browsing Video	606
14.5.1	Video Abstracts	606
14.5.2	Static Summaries	607
14.5.3	Mosaics and Salient Stills	608
14.5.4	Dynamic Summaries	609
14.5.5	Interactive Summaries	611
14.5.6	Visual vs. Audio Browsing	612
14.5.7	Evaluating Summaries	613

14.6	Fusion Models: Combining it All	614
14.6.1	Naming Faces	614
14.6.2	Naming Images	615
14.6.3	Naming Audio	616
14.6.4	Combining Audio and Video for AVSR	617
14.6.5	Combining Audio and Video for Multimedia	620
14.7	Segmentation	620
14.7.1	A Video Segmentation Example	620
14.7.2	Segmentation Schemes for Video	622
14.7.3	Video Segmentation with Edges	623
14.7.4	Speech Segmentation	624
14.7.5	Segmentation Evaluation	625
14.8	Compression and MPEG Standards	625
14.8.1	Intensity and Sampling	626
14.8.2	Color	626
14.8.3	Lossy Compression	628
14.8.4	Lossless Compression	628
14.8.5	Temporal Redundancy	630
14.8.6	Motion Prediction	631
14.8.7	MPEG Standards	633
14.9	Trends and Research Issues	636
14.10	Bibliographic Discussion	637
15	Enterprise Search	641
	<i>by David Hawking</i>	
15.1	Introduction	641
15.1.1	Characteristics and Applications of Enterprise Search	642
15.1.2	Enterprise Search Software	643
15.1.3	Workplace Search	644
15.2	Enterprise Search Tasks	644
15.2.1	Examples of Search-Supported Tasks	644
15.2.2	Search Types	647
15.2.3	Studying Enterprise Search	647
15.3	Architecture of Enterprise Search Systems	648
15.3.1	Gathering	648
15.3.2	Extracting	651
15.3.3	Indexing	652
15.3.4	Indexing Textual Annotations	653
15.3.5	Query Processing	654
15.3.6	Presentation of Search Results	655
15.3.7	Security Models	657
15.3.8	Federation/Metasearch	659
15.4	Enterprise Search Evaluation	662
15.4.1	Published Test Collections for Enterprise Search	662
15.4.2	Internal Enterprise Search Evaluations	663
15.4.3	Enterprise Search Tuning	665
15.4.4	What is it Reasonable to Expect?	666
15.5	Potential Reasons for Dissatisfaction	667

15.6	Context and Personalization	668
15.6.1	Controls and Levers for Contextualization	671
15.6.2	Contextualization: Local, Enterprise or Global?	675
15.6.3	Privacy of Profiles	676
15.6.4	Defining, Creating and Maintaining a Profile	677
15.6.5	User Modeling	677
15.6.6	Implicit Measures	679
15.6.7	Information Filtering	679
15.6.8	Social Recommender Systems	680
15.7	Trends and Research Issues	681
15.8	Bibliographic Discussion	681
16	Library Systems	685
	<i>by Edie Rasmussen</i>	
16.1	The Information Environment in the Library	685
16.2	Online Public Access Catalogues	687
16.2.1	OPACs and Bibliographic Records	689
16.2.2	Information Retrieval from the ILS	691
16.2.3	Integrating the Hybrid Library	693
16.2.4	OPACs and End Users	694
16.2.5	ILS: Vendors and Products	695
16.3	IR Systems and Document Databases	697
16.3.1	Bibliographic and Full-text Databases	698
16.3.2	Content of Database Records	698
16.3.3	The Online Industry: Database Vendors	701
16.3.4	Information Retrieval from Document Databases	702
16.4	Information Retrieval in Organizations	706
16.5	Trends and Research Issues	708
16.6	Bibliographic Discussion	709
17	Digital Libraries	711
	<i>by Marcos Gonçalves</i>	
17.1	Introduction	711
17.2	Defining Digital Libraries	712
17.3	A General Architecture	713
17.4	Fundamentals	714
17.4.1	Digital Objects and Collections	714
17.4.2	Metadata and Catalogs	716
17.4.3	Repositories/Archives	719
17.4.4	Services	723
17.5	Social-Economical Issues	725
17.5.1	Social Issues	725
17.5.2	Economical Issues	726
17.6	Software Systems	727
17.6.1	Greenstone	728
17.6.2	Eprints	728
17.6.3	DSpace	728
17.6.4	Fedora	729

17.6.5	Open Digital Libraries	729
17.6.6	The 5S Suite	730
17.7	DL Case Studies	731
17.7.1	The Networked DL of Theses and Dissertations	731
17.7.2	The National Science Digital Library	732
17.7.3	The ETANA-DL Archaeological Digital Library	732
17.8	Trends and Research Issues	733
17.8.1	Evaluation	733
17.8.2	Integration	733
17.8.3	Other Research Challenges	734
17.9	Bibliographic Discussion	735
A	Open Source Search Engines	737
	<i>with Christian Middleton</i>	
A.1	Introduction	737
A.2	Search Engines	738
A.2.1	Preliminary Selection of Search Engines	738
A.2.2	Features	741
A.2.3	Evaluation	742
A.3	Methodology	743
A.3.1	Document Collections	743
A.3.2	Evaluation Tests	744
A.3.3	Experimental Setup	744
A.4	Experimental Results	745
A.4.1	Test A – Indexing	745
A.4.2	Test B – Incremental Indexing	749
A.4.3	Test C – Search Performance	749
A.4.4	Global Evaluation	752
A.5	Conclusions	753
B	Biographies	755
	References	761
	Index	893