Modern Information Retrieval

the concepts and technology behind search

Second edition

Ricardo Baeza-Yates
Berthier Ribeiro-Neto

Modern Information Retrieval

The Concepts and Technology behind Search

Ricardo Baeza-Yates
Berthier Ribeiro-Neto

Second edition

Addison-Wesley

Harlow, England • Reading, Massachusetts
Menlo Park, California • New York
Don Mills, Ontario • Amsterdam • Bonn
Sydney • Singapore • Tokyo • Madrid
San Juan • Milan • Mexico City • Seoul • Taipei

To be filled by Pearson

To Helena, Rosa, and our children

Amo los libros exploradores,
libros con bosque o nieve,
profundidad o cielo

Un libro, un libro lleno
de contactos humanos, de camisas,
un libro sin soledad,
con hombres y herramientas,
un libro es la victoria.

de “Oda al Libro” (I) y (II),
en *Odas Elementales*, 1954.

Pablo Neruda

I love books that explore,
books with a forest or snow,
depth or sky

A book, a book full
of human contacts, of shirts,
a book without solitude,
with people and tools,
a book is the victory.

from “Ode to the Book” (I) and (II),
in *Elemental Odes*, 1954.

Pablo Neruda

território de homens livres
que será nosso país
e será pátria de todos.
Irmãos, cantai esse mundo
que não verei, mas virá
um dia, dentro de mil anos,
talvez mais... não tenho pressa.

de “Cidade Prevista” no livro
A Rosa do Povo, 1945.

Carlos Drummond de Andrade

territory of free men
that will be our country
and will be the nation of all
Brothers, sing that world
which I'll not see, but which will come
one day, in a thousand years,
maybe more... no hurry.

from “Prevised City” in the book
The Rose of the People, 1945.

Carlos Drummond de Andrade

Contents

Preface to the Second Edition xix

Preface to the First Edition xxi

Author's Acknowledgements to the Second Edition xxiii

Author's Acknowledgements to the First Edition xxv

Publisher's 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 9
  - 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 17
- 1.7 Bibliographic Discussion 17

2 User Interfaces for Search 21
by Marti Hearst
- 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 25
2.3 Search Interfaces Today 25
2.3.1 Getting Started 26
2.3.2 Query Specification 26
2.3.3 Query Specification Interfaces 27
2.3.4 Retrieval Results Display 30
2.3.5 Query Reformulation 32
2.3.6 Organizing Search Results 35
2.4 Visualization in Search Interfaces 42
2.4.1 Visualizing Boolean Syntax 43
2.4.2 Visualizing Query Terms within Retrieval Results 44
2.4.3 Visualizing Relationships Among Words and Documents 47
2.4.4 Visualization for Text Mining 48
2.5 Design and Evaluation of Search Interfaces 50
2.6 Trends and Research Issues 55
2.7 Bibliographic Discussion 55

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 62
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 76
3.2.6 The Vector Model 77
3.2.7 The Probabilistic Model 80
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 99
3.4.1 Generalized Vector Space Model 99
3.4.2 Latent Semantic Indexing Model 101
3.4.3 Neural Network Model 103
3.5 Alternative Probabilistic Models 105
3.5.1 BM25 105
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 126
3.6.3 Structured Text Retrieval 126
3.6.4 Multimedia Retrieval 126
3.6.5 Enterprise and Vertical Search 127
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 135
4.3.1 Precision and Recall 135
4.3.2 Single Value Summaries: P@n, MAP, MRR, F 140
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 158
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 170
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 175

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
with Gonzalo Navarro and Nivio Ziviani
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 209
6.4.1 SGML 209
6.4.2 HTML 210
6.4.3 XML 214
6.4.4 RDF: Resource Description Framework 217
6.4.5 HyTime 218
6.5 Text Properties 218
6.5.1 Information Theory 219
6.5.2 Modeling Natural Language 219
6.5.3 Text Similarity 222
6.6 Document Preprocessing 224
6.6.1 Lexical Analysis of the Text 225
6.6.2 Elimination of Stopwords 226
6.6.3 Stemming 226
6.6.4 Keyword Selection 228
6.6.5 Thesauri 228
6.7 Organizing Documents 231
6.7.1 Taxonomies 232
6.7.2 Folksonomies 232
6.8 Text Compression 233
6.8.1 Basic Concepts 234
6.8.2 Statistical Methods 235
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 247
6.8.7 Comparing Text Compression Techniques 248
6.8.8 Structured Text Compression 250
6.9 Trends and Research Issues 251
6.10 Bibliographical Discussion 254

7 Queries: Languages & Properties 257
with Gonzalo Navarro
7.1 Query Languages 257
7.1.1 Keyword-Based Querying 258
7.1.2 Beyond Keywords 261
7.1.3 Structural Queries 264
7.1.4 Query Protocols 267
7.2 Query Properties 269
7.2.1 Characterizing Web Queries 269
7.2.2 User Search Behavior 271
7.2.3 Query Intent 272
7.2.4 Query Topic 274
7.2.5 Query Sessions and Missions 275
7.2.6 Query Difficulty 276
7.3 Trends and Research Issues 279
7.4 Bibliographical Discussion 281

8 Text Classification 283
with Marcos Gonçalves
8.1 Introduction 283
8.2 A Characterization of Text Classification 284
8.2.1 Machine Learning 284
8.2.2 The Text Classification Problem 285
8.2.3 Text Classification Algorithms 286
8.3 Unsupervised Algorithms 287
8.3.1 Clustering 287
8.3.2 Naive Text Classification 292
8.4 Supervised Algorithms 293
8.4.1 Decision Trees 296
8.4.2 The k-NN Classifier 301
8.4.3 The Rocchio Classifier 302
8.4.4 Probabilistic Naive Bayes Document Classification 305
8.4.5 The SVM Classifier 309
8.4.6 Ensemble Classifiers 318
8.4.7 Final Remarks on Supervised Algorithms 321
8.5 Feature Selection or Dimensionality Reduction 322
8.5.1 Term–Class Incidence Table 323
8.5.2 Term Document Frequency 324
8.5.3 TF-IDF Weights 324
8.5.4 Mutual Information 325
8.5.5 Information Gain 325
8.5.6 Chi Square 326
8.5.7 Impact of Feature Selection 327
8.6 Evaluation Metrics 327
8.6.1 Contingency Table 327
8.6.2 Accuracy and Error 328
8.6.3 Precision and Recall 329

8.6.4 F-measure and F_1 . . . 329
8.6.5 Cross-Validation . . . 330
8.6.6 Standard Collections . . . 331
8.7 Organizing the Classes – Building Taxonomies . . . 332
8.8 Trends and Research Issues . . . 334
8.9 Bibliographic Discussion . . . 335

9 Indexing and Searching 339
with Gonzalo Navarro
9.1 Introduction . . . 339
9.2 Inverted Indexes . . . 342
9.2.1 Basic Concepts . . . 342
9.2.2 Full Inverted Indexes . . . 343
9.2.3 Searching . . . 347
9.2.4 Ranking . . . 351
9.2.5 Construction . . . 353
9.2.6 Compressed Inverted Indexes . . . 356
9.2.7 Structural Queries . . . 359
9.3 Signature Files . . . 359
9.4 Suffix Trees and Suffix Arrays . . . 362
9.4.1 Structure: Tries and Suffix Trees . . . 363
9.4.2 Searching for Simple Strings . . . 365
9.4.3 Searching for Complex Patterns . . . 366
9.4.4 Construction . . . 368
9.4.5 Compressed Suffix Arrays . . . 370
9.5 Sequential Searching . . . 374
9.5.1 Simple Strings: Horspool . . . 375
9.5.2 Complex Patterns: Automata and Bit-Parallelism . . . 378
9.5.3 Faster Bit-Parallel Algorithms . . . 382
9.5.4 Regular Expressions . . . 384
9.5.5 Multiple Patterns . . . 386
9.5.6 Approximate Searching . . . 387
9.5.7 Searching Compressed Text . . . 391
9.6 Multi-dimensional Indexing . . . 393
9.7 Trends and Research Issues . . . 395
9.8 Bibliographic Discussion . . . 396

10 Parallel and Distributed IR 401
with Eric Brown
10.1 Introduction . . . 401
10.2 A Taxonomy of Distributed IR Systems . . . 404
10.3 Data Partitioning . . . 406
10.3.1 Collection Partitioning . . . 407
10.3.2 Collection Selection . . . 409
10.3.3 Inverted Index Partitioning . . . 411
10.3.4 Partitioning other Indexes . . . 415
10.4 Parallel IR . . . 416

10.4.1 Introduction 416
10.4.2 Parallel IR on MIMD Architectures 419
10.4.3 Parallel IR on SIMD Architectures 420
10.5 Cluster-based IR 425
10.6 Distributed IR 426
10.6.1 Introduction 426
10.6.2 Indexing 431
10.6.3 Query Processing 433
10.6.4 Web Issues 439
10.7 Federated Search 440
10.8 Retrieval in Peer-to-Peer Networks 442
10.9 Trends and Research Issues 446
10.10Bibliographic Discussion 447

11 Web Retrieval 449
with Yoelle Maarek
11.1 Introduction 449
11.2 A Challenging Problem 451
11.3 The Web 453
11.3.1 Characteristics 453
11.3.2 Structure of the Web Graph 454
11.3.3 Modeling the Web 456
11.3.4 Link Analysis 458
11.4 Search Engine Architectures 460
11.4.1 Basic Architecture 460
11.4.2 Cluster-based Architecture 462
11.4.3 Caching 465
11.4.4 Multiple Indexes 467
11.4.5 Distributed Architectures 468
11.5 Search Engine Ranking 471
11.5.1 Ranking Signals 471
11.5.2 Link-based Ranking 472
11.5.3 Simple Ranking Functions 475
11.5.4 Learning to Rank 476
11.5.5 Learning the Ranking Function 477
11.5.6 Quality Evaluation 477
11.5.7 Web Spam 479
11.6 Managing Web Data 480
11.6.1 Assigning Identifiers to Documents 480
11.6.2 Metadata 480
11.6.3 Compressing the Web Graph 481
11.6.4 Handling Duplicated Data 481
11.7 Search Engine User Interaction 482
11.7.1 The Search Rectangle Paradigm 483
11.7.2 The Search Engine Result Page 491
11.7.3 Educating the User 500
11.8 Browsing 501

11.8.1 Flat Browsing 502
11.8.2 Structure Guided Browsing and Web Directories 502
11.9 Beyond Browsing 504
11.9.1 Hypertext and the Web 504
11.9.2 Combining Searching with Browsing 504
11.9.3 Web Query Languages 506
11.9.4 Dynamic Search 506
11.10Related Problems 507
11.10.1Computational Advertising 507
11.10.2Web Mining 509
11.10.3Metasearch 511
11.11Trends and Research Issues 512
11.11.1Beyond Static Text Data 513
11.11.2Current Challenges 514
11.12Bibliographical Discussion 516

12 Web Crawling 519
with Carlos Castillo
12.1 Introduction 519
12.2 Applications of a Web Crawler 521
12.2.1 General Web Search 521
12.2.2 Topical Crawling 522
12.2.3 Web Characterization 522
12.2.4 Mirroring 522
12.2.5 Web Site Analysis 523
12.3 A Taxonomy of Crawlers 523
12.3.1 Types of Web Pages 524
12.4 Architecture and Implementation 525
12.4.1 Crawler Architecture 525
12.4.2 Practical Issues 527
12.4.3 Parallel Crawling 530
12.5 Scheduling Algorithms 531
12.5.1 Selection Policy 532
12.5.2 Revisit Policy 535
12.5.3 Politeness Policy 540
12.5.4 Combining Policies 542
12.6 Evaluation 543
12.6.1 Evaluating Network Usage 543
12.6.2 Evaluating Long-term Scheduling 544
12.7 Trends and Research Issues 545
12.7.1 Crawling the “Hidden” Web 545
12.7.2 Crawling with the Help of Web Sites 546
12.7.3 Distributed Crawling 547
12.8 Bibliographic Discussion 547

13 Structured Text Retrieval 549
with Mounia Lalmas
13.1 Introduction 549
13.2 Structuring Power 550
13.2.1 Explicit vs. Implicit Structure 551
13.2.2 Static vs. Dynamic Structure 551
13.2.3 Single Hierarchy vs. Multiple Hierarchies 552
13.3 Early Text Retrieval Models 553
13.3.1 Model Based on Non-Overlapping Lists 553
13.3.2 Model Based on Proximal Nodes 554
13.3.3 Ranking Structured Text Results 555
13.4 XML Retrieval 555
13.4.1 Challenges in XML Retrieval 556
13.4.2 Indexing Strategies 557
13.4.3 Ranking Strategies 558
13.4.4 Removing Overlaps 569
13.5 XML Retrieval Evaluation 570
13.5.1 Document Collections 571
13.5.2 Topics 571
13.5.3 Retrieval Tasks 572
13.5.4 Relevance 574
13.5.5 Measures 576
13.6 Query Languages 578
13.6.1 Characteristics 578
13.6.2 Classification of XML Query Languages 579
13.6.3 Examples of XML Query Languages 581
13.7 Trends and Research Issues 586
13.8 Bibliographic Discussion 589

14 Multimedia Information Retrieval 591
by Dulce Ponceleón and Malcolm Slaney
14.1 Introduction 591
14.1.1 What is Multimedia? 591
14.1.2 Multimedia IR 592
14.1.3 Text IR versus Multimedia IR 593
14.2 The Challenges 594
14.2.1 The Semantic Gap 594
14.2.2 Feature Ambiguity 595
14.2.3 Machine-generated Data 595
14.3 Content-based Image Retrieval 596
14.3.1 Color-Based Retrieval 597
14.3.2 Texture 598
14.3.3 Salient Points 600
14.4 Audio and Music Retrieval 602
14.4.1 Fingerprinting 602
14.4.2 Speech Recognition 603
14.4.3 Speaker Identification 605

14.4.4 Spoken Document Retrieval 606
14.4.5 Audio Basics 606
14.5 Retrieving and Browsing Video 610
14.5.1 Video Abstracts 610
14.5.2 Static Summaries 611
14.5.3 Mosaics and Salient Stills 613
14.5.4 Dynamic Summaries 613
14.5.5 Interactive Summaries 615
14.5.6 Visual vs. Audio Browsing 617
14.5.7 Evaluating Summaries 618
14.6 Fusion Models: Combining it All 618
14.6.1 Naming Faces 618
14.6.2 Naming Images 620
14.6.3 Naming Audio 621
14.6.4 Combining Audio and Video for AVSR 621
14.6.5 Combining Audio and Video for Multimedia 624
14.7 Segmentation 624
14.7.1 A Video Segmentation Example 625
14.7.2 Segmentation Schemes for Video 626
14.7.3 Video Segmentation with Edges 627
14.7.4 Speech Segmentation 628
14.7.5 Segmentation Evaluation 629
14.8 Compression and MPEG Standards 630
14.8.1 Intensity and Sampling 630
14.8.2 Color 630
14.8.3 Lossy Compression 632
14.8.4 Lossless Compression 633
14.8.5 Temporal Redundancy 634
14.8.6 Motion Prediction 635
14.8.7 MPEG Standards 636
14.9 Trends and Research Issues 640
14.10Bibliographic Discussion 641

15 Enterprise Search **645**
by David Hawking
15.1 Introduction 645
15.1.1 Characteristics and Applications of Enterprise Search 646
15.1.2 Enterprise Search Software 647
15.1.3 Workplace Search 648
15.2 Enterprise Search Tasks 648
15.2.1 Examples of Search-Supported Tasks 648
15.2.2 Search Types 651
15.2.3 Studying Enterprise Search 651
15.3 Architecture of Enterprise Search Systems 652
15.3.1 Gathering 652
15.3.2 Extracting 654
15.3.3 Indexing 656

15.3.4 Indexing Textual Annotations 657
15.3.5 Query Processing 658
15.3.6 Presentation of Search Results 659
15.3.7 Security Models 661
15.3.8 Federation/Metasearch 663
15.4 Enterprise Search Evaluation 666
15.4.1 Published Test Collections for Enterprise Search 666
15.4.2 Internal Enterprise Search Evaluations 667
15.4.3 Enterprise Search Tuning 669
15.4.4 What is it Reasonable to Expect? 670
15.5 Potential Reasons for Dissatisfaction 671
15.6 Context and Personalization 672
15.6.1 Controls and Levers for Contextualization 675
15.6.2 Contextualization: Local, Enterprise or Global? 679
15.6.3 Privacy of Profiles 680
15.6.4 Defining, Creating and Maintaining a Profile 680
15.6.5 User Modeling 681
15.6.6 Implicit Measures 682
15.6.7 Information Filtering 683
15.6.8 Social Recommender Systems 684
15.7 Trends and Research Issues 684
15.8 Bibliographic Discussion 685

16 Library Systems 687
by Edie Rasmussen
16.1 The Information Environment in the Library 687
16.2 Online Public Access Catalogues 689
16.2.1 OPACs and Bibliographic Records 691
16.2.2 Information Retrieval from the ILS 693
16.2.3 Integrating the Hybrid Library 695
16.2.4 OPACs and End Users 696
16.2.5 ILS: Vendors and Products 697
16.3 IR Systems and Document Databases 699
16.3.1 Bibliographic and Full-text Databases 700
16.3.2 Content of Database Records 700
16.3.3 The Online Industry: Database Vendors 703
16.3.4 Information Retrieval from Document Databases 704
16.4 Information Retrieval in Organizations 708
16.5 Trends and Research Issues 710
16.6 Bibliographic Discussion 711

17 Digital Libraries 713
by Marcos Gonçalves
17.1 Introduction 713
17.2 Defining Digital Libraries 714
17.3 A General Architecture 715
17.4 Fundamentals 716

17.4.1 Digital Objects and Collections 716
17.4.2 Metadata and Catalogs 718
17.4.3 Repositories/Archives 721
17.4.4 Services 725
17.5 Social-Economical Issues 727
17.5.1 Social Issues 727
17.5.2 Economical Issues 728
17.6 Software Systems 729
17.6.1 Greenstone 730
17.6.2 Eprints 730
17.6.3 DSpace 730
17.6.4 Fedora 731
17.6.5 Open Digital Libraries 731
17.6.6 The 5S Suite 732
17.7 DL Case Studies 733
17.7.1 The Networked DL of Theses and Dissertations 733
17.7.2 The National Science Digital Library 734
17.7.3 The ETANA-DL Archaeological Digital Library 734
17.8 Trends and Research Issues 735
17.8.1 Evaluation 735
17.8.2 Integration 735
17.8.3 Other Research Challenges 736
17.9 Bibliographic Discussion 737

A Open Source Search Engines **739**
with Christian Middleton
A.1 Introduction 739
A.2 Search Engines 740
A.2.1 Preliminary Selection of Search Engines 740
A.2.2 Features 743
A.2.3 Evaluation 744
A.3 Methodology 745
A.3.1 Document Collections 745
A.3.2 Evaluation Tests 746
A.3.3 Experimental Setup 746
A.4 Experimental Results 747
A.4.1 Test A – Indexing 747
A.4.2 Test B – Incremental Indexing 751
A.4.3 Test C – Search Performance 751
A.4.4 Global Evaluation 754
A.5 Conclusions 755

B Biographies **757**

References **765**