

# Contents

---

Foreword	<i>xix</i>
Preface	<i>xxi</i>
Notation	<i>xxv</i>

---

## I Foundations 1

---

1	Introduction	2
1.1	What Is Information Retrieval?	2
1.1.1	Web Search	2
1.1.2	Other Search Applications	3
1.1.3	Other IR Applications	4
1.2	Information Retrieval Systems	5
1.2.1	Basic IR System Architecture	5
1.2.2	Documents and Update	7
1.2.3	Performance Evaluation	8
1.3	Working with Electronic Text	9
1.3.1	Text Formats	9
1.3.2	A Simple Tokenization of English Text	13
1.3.3	Term Distributions	15
1.3.4	Language Modeling	17
1.4	Test Collections	23
1.4.1	TREC Tasks	24
1.5	Open-Source IR Systems	27
1.5.1	Lucene	27
1.5.2	Indri	27
1.5.3	Wumpus	28

1.6	Further Reading	28
1.7	Exercises	30
1.8	Bibliography	32
<b>2</b>	<b>Basic Techniques</b>	<b>33</b>
2.1	Inverted Indices	33
2.1.1	Extended Example: Phrase Search	35
2.1.2	Implementing Inverted Indices	39
2.1.3	Documents and Other Elements	45
2.2	Retrieval and Ranking	51
2.2.1	The Vector Space Model	54
2.2.2	Proximity Ranking	60
2.2.3	Boolean Retrieval	63
2.3	Evaluation	66
2.3.1	Recall and Precision	67
2.3.2	Effectiveness Measures for Ranked Retrieval	68
2.3.3	Building a Test Collection	73
2.3.4	Efficiency Measures	75
2.4	Summary	76
2.5	Further Reading	77
2.6	Exercises	79
2.7	Bibliography	82
<b>3</b>	<b>Tokens and Terms</b>	<b>84</b>
3.1	English	85
3.1.1	Punctuation and Capitalization	85
3.1.2	Stemming	86
3.1.3	Stopping	89
3.2	Characters	91
3.3	Character N-Grams	92
3.4	European Languages	94
3.5	CJK Languages	95
3.6	Further Reading	97
3.7	Exercises	99
3.8	Bibliography	100

---

<b>II Indexing</b>	<b>103</b>
<b>4 Static Inverted Indices</b>	<b>104</b>
4.1 Index Components and Index Life Cycle	104
4.2 The Dictionary	106
4.3 Postings Lists	110
4.4 Interleaving Dictionary and Postings Lists	114
4.5 Index Construction	118
4.5.1 In-Memory Index Construction	119
4.5.2 Sort-Based Index Construction	125
4.5.3 Merge-Based Index Construction	127
4.6 Other Types of Indices	131
4.7 Summary	132
4.8 Further Reading	132
4.9 Exercises	133
4.10 Bibliography	135
<b>5 Query Processing</b>	<b>137</b>
5.1 Query Processing for Ranked Retrieval	137
5.1.1 Document-at-a-Time Query Processing	139
5.1.2 Term-at-a-Time Query Processing	145
5.1.3 Precomputing Score Contributions	151
5.1.4 Impact Ordering	153
5.1.5 Static Index Pruning	153
5.2 Lightweight Structure	160
5.2.1 Generalized Concordance Lists	160
5.2.2 Operators	162
5.2.3 Examples	163
5.2.4 Implementation	165
5.3 Further Reading	169
5.4 Exercises	170
5.5 Bibliography	171

## 6 Index Compression 174

- 6.1 General-Purpose Data Compression 175
- 6.2 Symbolwise Data Compression 176
  - 6.2.1 Modeling and Coding 177
  - 6.2.2 Huffman Coding 181
  - 6.2.3 Arithmetic Coding 186
  - 6.2.4 Symbolwise Text Compression 189
- 6.3 Compressing Postings Lists 191
  - 6.3.1 Nonparametric Gap Compression 192
  - 6.3.2 Parametric Gap Compression 195
  - 6.3.3 Context-Aware Compression Methods 201
  - 6.3.4 Index Compression for High Query Performance 204
  - 6.3.5 Compression Effectiveness 209
  - 6.3.6 Decoding Performance 212
  - 6.3.7 Document Reordering 214
- 6.4 Compressing the Dictionary 216
- 6.5 Summary 222
- 6.6 Further Reading 223
- 6.7 Exercises 224
- 6.8 Bibliography 225

## 7 Dynamic Inverted Indices 228

- 7.1 Batch Updates 229
- 7.2 Incremental Index Updates 231
  - 7.2.1 Contiguous Inverted Lists 233
  - 7.2.2 Noncontiguous Inverted Lists 239
- 7.3 Document Deletions 243
  - 7.3.1 Invalidation List 243
  - 7.3.2 Garbage Collection 245
- 7.4 Document Modifications 250
- 7.5 Discussion and Further Reading 251
- 7.6 Exercises 253
- 7.7 Bibliography 254

---

**III Retrieval and Ranking 257**


---

- 8 Probabilistic Retrieval 258**
- 8.1 Modeling Relevance 259
  - 8.2 The Binary Independence Model 261
  - 8.3 The Robertson/Spärck Jones Weighting Formula 264
  - 8.4 Term Frequency 266
    - 8.4.1 Bookstein's Two-Poisson Model 267
    - 8.4.2 Approximating the Two-Poisson Model 270
    - 8.4.3 Query Term Frequency 271
  - 8.5 Document Length: BM25 271
  - 8.6 Relevance Feedback 273
    - 8.6.1 Term Selection 274
    - 8.6.2 Pseudo-Relevance Feedback 275
  - 8.7 Field Weights: BM25F 277
  - 8.8 Experimental Comparison 279
  - 8.9 Further Reading 280
  - 8.10 Exercises 281
  - 8.11 Bibliography 282
- 9 Language Modeling and Related Methods 286**
- 9.1 Generating Queries from Documents 287
  - 9.2 Language Models and Smoothing 289
  - 9.3 Ranking with Language Models 292
  - 9.4 Kullback-Leibler Divergence 296
  - 9.5 Divergence from Randomness 298
    - 9.5.1 A Model of Randomness 299
    - 9.5.2 Eliteness 301
    - 9.5.3 Document Length Normalization 301
  - 9.6 Passage Retrieval and Ranking 302
    - 9.6.1 Passage Scoring 304
    - 9.6.2 Implementation 304
  - 9.7 Experimental Comparison 306
  - 9.8 Further Reading 306
- 
- IV Evaluation**
-

9.9	Exercises	307
9.10	Bibliography	307
<b>10</b>	<b>Categorization and Filtering</b>	<b>310</b>
10.1	Detailed Examples	313
10.1.1	Topic-Oriented Batch Filtering	313
10.1.2	On-Line Filtering	317
10.1.3	Learning from Historical Examples	318
10.1.4	Language Categorization	320
10.1.5	On-Line Adaptive Spam Filtering	325
10.1.6	Threshold Choice for Binary Categorization	329
10.2	Classification	331
10.2.1	Odds and Odds Ratios	333
10.2.2	Building Classifiers	334
10.2.3	Learning Modes	336
10.2.4	Feature Engineering	338
10.3	Probabilistic Classifiers	339
10.3.1	Probability Estimates	340
10.3.2	Combining Probability Estimates	343
10.3.3	Practical Considerations	347
10.4	Linear Classifiers	349
10.4.1	Perceptron Algorithm	352
10.4.2	Support Vector Machines	353
10.5	Similarity-Based Classifiers	354
10.5.1	Rocchio's Method	354
10.5.2	Memory-Based Methods	355
10.6	Generalized Linear Models	355
10.6.1	Kernel Methods	357
10.7	Information-Theoretic Models	359
10.7.1	Comparing Models	360
10.7.2	Sequential Compression Models	361
10.7.3	Decision Trees and Stumps	364
10.8	Experimental Comparison	366
10.8.1	Topic-Oriented On-Line Filtering	367
10.8.2	On-Line Adaptive Spam Filtering	369
10.9	Further Reading	371

10.10	Exercises	372
10.11	Bibliography	373
<b>11</b>	<b>Fusion and Metalearning</b>	<b>376</b>
11.1	Search-Result Fusion	377
11.1.1	Fixed-Cutoff Aggregation	379
11.1.2	Rank and Score Aggregation	380
11.2	Stacking Adaptive Filters	381
11.3	Stacking Batch Classifiers	383
11.3.1	Holdout Validation	383
11.3.2	Cross-Validation	384
11.4	Bagging	385
11.5	Boosting	387
11.6	Multicategory Ranking and Classification	388
11.6.1	Document Versus Category Scores	389
11.6.2	Document Versus Category Rank Fusion	390
11.6.3	Multicategory Methods	391
11.7	Learning to Rank	394
11.7.1	What Is Learning to Rank?	395
11.7.2	Learning-to-Rank Methods	396
11.7.3	What to Optimize?	396
11.7.4	Learning to Rank for Categorization	397
11.7.5	Learning for Ranked IR	398
11.7.6	The LETOR Data Set	399
11.8	Further Reading	400
11.9	Exercises	401
11.10	Bibliography	401

---

## IV Evaluation 405

---

<b>12</b>	<b>Measuring Effectiveness</b>	<b>406</b>
12.1	Traditional Effectiveness Measures	407
12.1.1	Recall and Precision	407
12.1.2	Precision at $k$ Documents ( $P@k$ )	408
12.1.3	Average Precision (AP)	408

12.1.4	Reciprocal Rank (RR)	409
12.1.5	Arithmetic Mean Versus Geometric Mean	409
12.1.6	User Satisfaction	410
12.2	The Text REtrieval Conference (TREC)	410
12.3	Using Statistics in Evaluation	412
12.3.1	Foundations and Terminology	413
12.3.2	Confidence Intervals	416
12.3.3	Comparative Evaluation	424
12.3.4	Hypothesis Tests Considered Harmful	427
12.3.5	Paired and Unpaired Differences	429
12.3.6	Significance Tests	430
12.3.7	Validity and Power of Statistical Tests	434
12.3.8	Reporting the Precision of Measurement	438
12.3.9	Meta-Analysis	439
12.4	Minimizing Adjudication Effort	441
12.4.1	Selecting Documents for Adjudication	443
12.4.2	Sampling the Pool	449
12.5	Nontraditional Effectiveness Measures	451
12.5.1	Graded Relevance	451
12.5.2	Incomplete and Biased Judgments	453
12.5.3	Novelty and Diversity	455
12.6	Further Reading	460
12.7	Exercises	462
12.8	Bibliography	463
<b>13</b>	<b>Measuring Efficiency</b>	<b>468</b>
13.1	Efficiency Criteria	468
13.1.1	Throughput and Latency	469
13.1.2	Aggregate Statistics and User Satisfaction	472
13.2	Queueing Theory	472
13.2.1	Kendall's Notation	474
13.2.2	The M/M/1 Queueing Model	475
13.2.3	Latency Quantiles and Average Utilization	477
13.3	Query Scheduling	478
13.4	Caching	479
13.4.1	Three-Level Caching	480



- 13.4.2 Cache Policies 482
- 13.4.3 Prefetching Search Results 483
- 13.5 Further Reading 484
- 13.6 Exercises 484
- 13.7 Bibliography 485

---

## V Applications and Extensions 487

---

### 14 Parallel Information Retrieval 488

- 14.1 Parallel Query Processing 488
  - 14.1.1 Document Partitioning 490
  - 14.1.2 Term Partitioning 493
  - 14.1.3 Hybrid Schemes 495
  - 14.1.4 Redundancy and Fault Tolerance 496
- 14.2 MapReduce 498
  - 14.2.1 The Basic Framework 498
  - 14.2.2 Combiners 500
  - 14.2.3 Secondary Keys 502
  - 14.2.4 Machine Failures 502
- 14.3 Further Reading 503
- 14.4 Exercises 504
- 14.5 Bibliography 505

### 15 Web Search 507

- 15.1 The Structure of the Web 508
  - 15.1.1 The Web Graph 508
  - 15.1.2 Static and Dynamic Pages 510
  - 15.1.3 The Hidden Web 511
  - 15.1.4 The Size of the Web 511
- 15.2 Queries and Users 513
  - 15.2.1 User Intent 513
  - 15.2.2 Clickthrough Curves 516
- 15.3 Static Ranking 517
  - 15.3.1 Basic PageRank 517
  - 15.3.2 Extended PageRank 522

15.3.3	Properties of PageRank	528
15.3.4	Other Link Analysis Methods: HITS and SALSA	532
15.3.5	Other Static Ranking Methods	535
15.4	Dynamic Ranking	535
15.4.1	Anchor Text	536
15.4.2	Novelty	537
<hr/>		
15.5	Evaluating Web Search	538
15.5.1	Named Page Finding	538
15.5.2	Implicit User Feedback	540
15.6	Web Crawlers	541
15.6.1	Components of a Crawler	542
15.6.2	Crawl Order	547
15.6.3	Duplicates and Near-Duplicates	549
15.7	Summary	553
15.8	Further Reading	553
15.8.1	Link Analysis	554
15.8.2	Anchor Text	555
15.8.3	Implicit Feedback	555
15.8.4	Web Crawlers	556
15.9	Exercises	556
15.10	Bibliography	558
<b>16</b>	<b>XML Retrieval</b>	<b>564</b>
16.1	The Essence of XML	565
16.1.1	Document Type Definitions	568
16.1.2	XML Schema	570
16.2	Paths, Trees, and FLWORS	571
16.2.1	XPath	571
16.2.2	NEXI	572
16.2.3	XQuery	574
16.3	Indexing and Query Processing	576
16.4	Ranked Retrieval	579
16.4.1	Ranking Elements	580
16.4.2	Overlapping Elements	582
16.4.3	Retrievable Elements	583
16.5	Evaluation	583

## Foreword

16.5.1	Test Collections	583
16.5.2	Effectiveness Measures	584
16.6	Further Reading	585
16.7	Exercises	587
16.8	Bibliography	587

---

**VI Appendix 591**

<b>A</b>	<b>Computer Performance</b>	<b>592</b>
A.1	Sequential Versus Random Access on Disk	592
A.2	Sequential Versus Random Access in RAM	593
A.3	Pipelined Execution and Branch Prediction	594
<b>Index</b>	<b>597</b>	

The authors provide a tutorial overview of current information retrieval research, with hundreds of references into the research literature, but they go well beyond the typical survey. Using a running set of examples and a common framework, they describe in concrete terms the important methods underlying each component — why they work, how they may be implemented, and how they may be shown to work. For the purpose of this book, the authors have implemented and tested nearly every important method, conducting hundreds of experiments whose results augment the exposition. Exercises at the end of each chapter encourage you to build and explore on your own.

This book is a must-read for all search academics and practitioners!

Amit Singhal, Google Fellow