

Contents

<i>Table of Notation</i>	page xi
<i>Preface</i>	xv
1 Boolean retrieval	1
1.1 An example information retrieval problem	3
1.2 A first take at building an inverted index	6
1.3 Processing Boolean queries	9
1.4 The extended Boolean model versus ranked retrieval	13
1.5 References and further reading	16
2 The term vocabulary and postings lists	18
2.1 Document delineation and character sequence decoding	18
2.2 Determining the vocabulary of terms	21
2.3 Faster postings list intersection via skip pointers	33
2.4 Positional postings and phrase queries	36
2.5 References and further reading	43
3 Dictionaries and tolerant retrieval	45
3.1 Search structures for dictionaries	45
3.2 Wildcard queries	48
3.3 Spelling correction	52
3.4 Phonetic correction	58
3.5 References and further reading	59
4 Index construction	61
4.1 Hardware basics	62
4.2 Blocked sort-based indexing	63
4.3 Single-pass in-memory indexing	66
4.4 Distributed indexing	68
4.5 Dynamic indexing	71

4.6 Other types of indexes	73
4.7 References and further reading	76
5 Index compression	78
5.1 Statistical properties of terms in information retrieval	79
5.2 Dictionary compression	82
5.3 Postings file compression	87
5.4 References and further reading	97
6 Scoring, term weighting, and the vector space model	100
6.1 Parametric and zone indexes	101
6.2 Term frequency and weighting	107
6.3 The vector space model for scoring	110
6.4 Variant tf-idf functions	116
6.5 References and further reading	122
7 Computing scores in a complete search system	124
7.1 Efficient scoring and ranking	124
7.2 Components of an information retrieval system	132
7.3 Vector space scoring and query operator interaction	136
7.4 References and further reading	137
8 Evaluation in information retrieval	139
8.1 Information retrieval system evaluation	140
8.2 Standard test collections	141
8.3 Evaluation of unranked retrieval sets	142
8.4 Evaluation of ranked retrieval results	145
8.5 Assessing relevance	151
8.6 A broader perspective: System quality and user utility	154
8.7 Results snippets	157
8.8 References and further reading	159
9 Relevance feedback and query expansion	162
9.1 Relevance feedback and pseudo relevance feedback	163
9.2 Global methods for query reformulation	173
9.3 References and further reading	177
10 XML retrieval	178
10.1 Basic XML concepts	180
10.2 Challenges in XML retrieval	183
10.3 A vector space model for XML retrieval	188
10.4 Evaluation of XML retrieval	192

10.5 Text-centric versus data-centric XML retrieval	196
10.6 References and further reading	198
11 Probabilistic information retrieval	201
11.1 Review of basic probability theory	202
11.2 The probability ranking principle	203
11.3 The binary independence model	204
11.4 An appraisal and some extensions	212
11.5 References and further reading	216
12 Language models for information retrieval	218
12.1 Language models	218
12.2 The query likelihood model	223
12.3 Language modeling versus other approaches in information retrieval	229
12.4 Extended language modeling approaches	230
12.5 References and further reading	232
13 Text classification and Naive Bayes	234
13.1 The text classification problem	237
13.2 Naive Bayes text classification	238
13.3 The Bernoulli model	243
13.4 Properties of Naive Bayes	245
13.5 Feature selection	251
13.6 Evaluation of text classification	258
13.7 References and further reading	264
14 Vector space classification	266
14.1 Document representations and measures of relatedness in vector spaces	267
14.2 Rocchio classification	269
14.3 k nearest neighbor	273
14.4 Linear versus nonlinear classifiers	277
14.5 Classification with more than two classes	281
14.6 The bias-variance tradeoff	284
14.7 References and further reading	291
15 Support vector machines and machine learning on documents	293
15.1 Support vector machines: The linearly separable case	294
15.2 Extensions to the support vector machine model	300
15.3 Issues in the classification of text documents	307
15.4 Machine-learning methods in ad hoc information retrieval	314
15.5 References and further reading	318

16 Flat clustering	321
16.1 Clustering in information retrieval	322
16.2 Problem statement	326
16.3 Evaluation of clustering	327
16.4 K-means	331
16.5 Model-based clustering	338
16.6 References and further reading	343
17 Hierarchical clustering	346
17.1 Hierarchical agglomerative clustering	347
17.2 Single-link and complete-link clustering	350
17.3 Group-average agglomerative clustering	356
17.4 Centroid clustering	358
17.5 Optimality of hierarchical agglomerative clustering	360
17.6 Divisive clustering	362
17.7 Cluster labeling	363
17.8 Implementation notes	365
17.9 References and further reading	367
18 Matrix decompositions and latent semantic indexing	369
18.1 Linear algebra review	369
18.2 Term-document matrices and singular value decompositions	373
18.3 Low-rank approximations	376
18.4 Latent semantic indexing	378
18.5 References and further reading	383
19 Web search basics	385
19.1 Background and history	385
19.2 Web characteristics	387
19.3 Advertising as the economic model	392
19.4 The search user experience	395
19.5 Index size and estimation	396
19.6 Near-duplicates and shingling	400
19.7 References and further reading	404
20 Web crawling and indexes	405
20.1 Overview	405
20.2 Crawling	406
20.3 Distributing indexes	415
20.4 Connectivity servers	416
20.5 References and further reading	419

<i>Contents</i>	ix
21 <i>Link analysis</i>	421
21.1 The Web as a graph	422
21.2 PageRank	424
21.3 Hubs and authorities	433
21.4 References and further reading	439
<i>Bibliography</i>	441
<i>Index</i>	469