Contents

Table of Notation xi
Preface 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
 Contents

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
Contents

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

21 Link analysis 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

Bibliography 441
Index 469