

Contents

Foreword	xiii
Preface	xv
About the Editors	xvii
List of Contributors	xix
Introduction	xxi
1 Information Retrieval Models	1
<i>Djoerd Hiemstra</i>	
1.1 Introduction	1
1.1.1 Terminology	1
1.1.2 What is a model?	2
1.1.3 Outline	3
1.2 Exact Match Models	3
1.2.1 The Boolean model	3
1.2.2 Region models	4
1.2.3 Discussion	5
1.3 Vector Space Approaches	5
1.3.1 The vector space model	6
1.3.2 Positioning the query in vector space	7
1.3.3 Term weighting and other caveats	7
1.4 Probabilistic Approaches	8
1.4.1 The probabilistic indexing model	8
1.4.2 The probabilistic retrieval model	9
1.4.3 The 2-Poisson model	10
1.4.4 Bayesian network models	11
1.4.5 Language models	12
1.4.6 Google's PageRank model	14
1.5 Summary and Further Reading	15
Exercises	16
References	18
2 User-centred Evaluation of Information Retrieval Systems	21
<i>Pia Borlund</i>	
2.1 Introduction	21
2.1.1 Background	21
2.1.2 Chapter outline	23
2.2 The MEDLARS Test	23

2.2.1	<i>Description of Lancaster's test of MEDLARS</i>	24
2.2.2	<i>Evaluation characteristics</i>	25
2.3	The Okapi Project	26
2.3.1	<i>The objectives of Okapi</i>	26
2.3.2	<i>Okapi at TREC</i>	27
2.3.3	<i>The impact of Okapi</i>	28
2.4	The Interactive IR Evaluation Model	28
2.4.1	<i>The cognitive IR approach</i>	29
2.4.2	<i>The three parts of the IIR evaluation model</i>	29
2.5	Summary	31
	Exercises	34
	References	35
3	Multimedia Resource Discovery	39
	<i>Stefan Ruger</i>	
3.1	Introduction	39
3.2	Basic Multimedia Search Technologies	40
3.2.1	<i>Piggy-back text retrieval</i>	41
3.2.2	<i>Automated annotation</i>	41
3.2.3	<i>Content-based retrieval</i>	42
3.2.4	<i>Fingerprinting</i>	44
3.3	Challenges of Automated Visual Indexing	44
3.4	Added Services	45
3.4.1	<i>Video summaries</i>	45
3.4.2	<i>New paradigms in information visualisation</i>	46
3.4.3	<i>Visual search and relevance feedback</i>	49
3.5	Browsing: Lateral and Geotemporal	51
3.6	Summary	55
	Exercises	57
	References	59
4	Image Users' Needs and Searching Behaviour	63
	<i>Stina Westman</i>	
4.1	Introduction	63
4.2	Image Attributes and Users' Needs	64
4.2.1	<i>Image attributes</i>	64
4.2.2	<i>Image attributes in queries</i>	67
4.2.3	<i>Attributes beyond queries</i>	69
4.2.4	<i>Image needs</i>	70
4.3	Image Searching Behaviour	71
4.3.1	<i>Search process</i>	71
4.3.2	<i>Search strategies</i>	73
4.3.3	<i>Relevance criteria</i>	76
4.4	New Directions for Image Access	77
4.4.1	<i>Social tagging</i>	77
4.4.2	<i>Images in context</i>	77
4.4.3	<i>Visualisations</i>	78
4.4.4	<i>Workspaces</i>	78
4.5	Summary	78
	Exercise	80
	References	81

5	Web Information Retrieval	85
	<i>Nick Craswell and David Hawking</i>	
5.1	Introduction	85
5.2	Distinctive Characteristics of the Web	85
	5.2.1 <i>Web data</i>	85
	5.2.2 <i>Web structure</i>	86
	5.2.3 <i>User behaviour</i>	87
	5.2.4 <i>User interaction data</i>	87
5.3	Three Ranking Problems	87
	5.3.1 <i>Retrieval</i>	88
	5.3.2 <i>Selective crawling</i>	88
	5.3.3 <i>Index organisation</i>	89
5.4	Other Web IR Issues	90
	5.4.1 <i>Stemming</i>	90
	5.4.2 <i>Treatment of near-duplicate content</i>	90
	5.4.3 <i>Spelling suggestions</i>	91
	5.4.4 <i>Spam rejection</i>	91
	5.4.5 <i>Adult content filtering – genre classification</i>	92
	5.4.6 <i>Query-targeted advertisement generation</i>	92
	5.4.7 <i>Snippet generation</i>	92
	5.4.8 <i>Context and web information retrieval</i>	93
5.5	Evaluation of Web Search Effectiveness	93
	5.5.1 <i>TREC-9 Web Track: Realistic queries, rich link structure, traditional IR task</i>	94
	5.5.2 <i>Evaluation using web-specific tasks</i>	95
	5.5.3 <i>Future directions for web IR evaluations</i>	95
	5.5.4 <i>Comparing results lists in context</i>	96
	5.5.5 <i>Evaluation by commercial web search companies</i>	96
5.6	Summary	97
	Exercises	98
	References	99
6	Mobile Search	103
	<i>David Mountain, Hans Myrhaug and Ayşe Göker</i>	
6.1	Introduction: Mobile Search – Why Now?	103
	6.1.1 <i>Technological drivers</i>	103
	6.1.2 <i>Predicted demand for mobile search</i>	104
6.2	Information for Mobile Search	107
	6.2.1 <i>Linking information to physical space</i>	107
	6.2.2 <i>The storage of information</i>	107
	6.2.3 <i>The ownership of information</i>	113
6.3	Designing for Mobile Search	115
	6.3.1 <i>Characteristics of mobile usage</i>	115
	6.3.2 <i>Filters as a framework for mobile search</i>	116
	6.3.3 <i>Manual versus automatic filtering</i>	119
	6.3.4 <i>Using filters to push information</i>	120
6.4	Case Studies	121
	6.4.1 <i>Oslo airport – AmbieSense</i>	121
	6.4.2 <i>The Swiss Alps – WebPark</i>	123
6.5	Summary	125
	Exercises	127
	References	129

7	Context and Information Retrieval	131
	<i>Ayşe Göker, Hans Myrhaug and Ralf Bierig</i>	
7.1	Introduction	131
7.2	What is Context?	132
	7.2.1 <i>Whose context?</i>	133
7.3	Context in Information Retrieval	133
	7.3.1 <i>Context in the wider sense</i>	134
	7.3.2 <i>Perceptions of context in related fields</i>	135
	7.3.3 <i>Example: context and images</i>	135
7.4	Context Modelling and Representation	136
	7.4.1 <i>Context modelling</i>	137
	7.4.2 <i>User models and their relationship to context</i>	140
	7.4.3 <i>Past, present and future contexts</i>	141
7.5	Context and Content	143
	7.5.1 <i>Representation of context</i>	143
	7.5.2 <i>Capturing context</i>	144
	7.5.3 <i>Searching with context information</i>	145
	7.5.4 <i>Context templates</i>	146
7.6	Related Topics	147
	7.6.1 <i>Personalisation and context</i>	147
	7.6.2 <i>Mobility and context</i>	148
7.7	Evaluating Context-aware IR Systems	148
	7.7.1 <i>Principles of methodology</i>	149
7.8	Summary	150
	Exercises	151
	References	153
8	Text Categorisation and Genre in Information Retrieval	159
	<i>Stuart Watt</i>	
8.1	Introduction: What is Text Categorisation?	159
	8.1.1 <i>Purpose of categorisation</i>	160
8.2	How to Build a Text Categorisation System	162
	8.2.1 <i>The classifier component</i>	163
	8.2.2 <i>The machine learning component</i>	166
	8.2.3 <i>The feature selection component</i>	167
8.3	Evaluating Text Categorisation Systems	168
8.4	Genre: Text Structure and Purpose	169
	8.4.1 <i>An overview of genre</i>	169
	8.4.2 <i>Text categorisation and genre</i>	171
	8.4.3 <i>The importance of layout</i>	171
8.5	Related Techniques: Information Filtering	172
8.6	Applications of Text Categorisation	173
8.7	Summary and the Future of Text Categorisation	174
	Exercises	175
	References	176
9	Semantic Search	179
	<i>John Davies, Alistair Duke and Atanas Kiryakov</i>	
9.1	Introduction	179
	9.1.1 <i>Limitations of current search technology</i>	180
	9.1.2 <i>Ontologies</i>	181
	9.1.3 <i>Knowledge bases and semantic repositories</i>	182

9.2	Semantic Web	184
9.2.1	<i>Semantic web and semantic search</i>	185
9.2.2	<i>Basic semantic web standards: RDF(S) and OWL</i>	185
9.3	Metadata and Annotations	187
9.4	Semantic Annotations: the Fibres of the Semantic Web	189
9.5	Semantic Annotation of Named Entities	191
9.5.1	<i>Named entities</i>	192
9.5.2	<i>Semantic annotation model and representation</i>	192
9.6	Semantic Indexing and Retrieval	193
9.6.1	<i>Indexing with respect to lexical concepts</i>	194
9.6.2	<i>Indexing with respect to named entities</i>	196
9.6.3	<i>Retrieval as spreading activation over semantic network</i>	196
9.7	Semantic Search Tools	197
9.7.1	<i>Searching through document-level RDF annotations – QuizRDF</i>	197
9.7.2	<i>Exploiting massive background knowledge – TAP</i>	200
9.7.3	<i>Character-level annotations and massive world knowledge – KIM</i>	200
9.7.4	<i>Squirrel</i>	203
9.7.5	<i>Other approaches</i>	207
9.8	Summary	208
	Exercises	210
	References	211
10	The Role of Natural Language Processing in Information Retrieval: Searching for Meaning and Structure	215
	<i>Tony Russell-Rose and Mark Stevenson</i>	
10.1	Introduction	215
10.2	Natural Language Processing Techniques	217
10.2.1	<i>Named entity recognition</i>	217
10.2.2	<i>Information extraction</i>	218
10.2.3	<i>WordNet</i>	220
10.2.4	<i>Word sense disambiguation</i>	221
10.2.5	<i>Evaluation</i>	222
10.3	Applications of Natural Language Processing in Information Retrieval	223
10.3.1	<i>Text mining</i>	223
10.3.2	<i>Question answering</i>	224
10.4	Discussion	225
10.5	Summary	226
	Exercises	228
	References	229
11	Cross-Language Information Retrieval	233
	<i>Daqing He and Jianqiang Wang</i>	
11.1	Introduction	233
11.2	Major Approaches and Challenges in CLIR	234
11.3	Identifying Translation Units	235
11.3.1	<i>Tokenisation</i>	236
11.3.2	<i>Stemming</i>	236
11.3.3	<i>Phrase identification</i>	236
11.3.4	<i>Stop-words</i>	237
11.4	Obtaining Translation Knowledge	237
11.4.1	<i>Obtaining bilingual dictionaries and corpora</i>	237
11.4.2	<i>Extracting translation knowledge</i>	238

11.4.3	<i>Dealing with out-of-vocabulary terms</i>	238
11.5	Using Translation Knowledge	239
11.5.1	<i>Translation disambiguation</i>	239
11.5.2	<i>Weighting translation alternatives</i>	240
11.5.3	<i>Using translation probabilities in term weighting</i>	241
11.6	Interactivity in CLIR	241
11.6.1	<i>Interactive CLIR</i>	241
11.6.2	<i>Query translation in interactive CLIR</i>	242
11.6.3	<i>Document selection in interactive CLIR</i>	243
11.7	Evaluation of CLIR Systems	244
11.7.1	<i>Cranfield-based evaluation framework</i>	244
11.7.2	<i>Evaluations on interactive CLIR</i>	245
11.7.3	<i>Current CLIR evaluation frameworks</i>	245
11.8	Summary and Future Directions	245
11.8.1	<i>Current achievements in CLIR</i>	245
11.8.2	<i>Future directions for CLIR</i>	246
11.8.3	<i>Further reading</i>	246
	Exercises	247
	References	248
12	Performance Issues in Parallel Computing for Information Retrieval	255
	<i>Andrew MacFarlane</i>	
12.1	Introduction	255
12.2	Why Parallel IR?	255
12.3	Review of Previous Work	257
12.4	Distribution Methods for Inverted File Data	257
12.4.1	<i>On-the-fly distribution</i>	258
12.4.2	<i>Inverted file replication</i>	259
12.4.3	<i>Inverted file partitioning</i>	259
12.5	Tasks in Information Retrieval	260
12.5.1	<i>The indexing task</i>	261
12.5.2	<i>The probabilistic search task</i>	261
12.5.3	<i>The passage retrieval task</i>	261
12.5.4	<i>The routing/filtering task</i>	262
12.5.5	<i>The index update task</i>	262
12.6	A Synthetic Model of Performance for Parallel Information Retrieval	262
12.7	Empirical Examination of Synthetic Model	264
12.7.1	<i>Comparative results using indexing models</i>	264
12.7.2	<i>Comparative results using search models</i>	265
12.7.3	<i>Comparative results using passage retrieval models</i>	265
12.7.4	<i>Comparative results using term selection models</i>	267
12.7.5	<i>Comparative results using index update model</i>	268
12.8	Summary and Further Research	269
	Exercises	270
	References	271
	Solutions to Exercises	273
	Index	285