

Contents

Part I Principles of Information Retrieval

1	An Introduction to Information Retrieval	3
1.1	What Is Information Retrieval?	3
1.1.1	Defining Relevance	4
1.1.2	Dealing with Large, Unstructured Data Collections	4
1.1.3	Formal Characterization	5
1.1.4	Typical Information Retrieval Tasks	5
1.2	Evaluating an Information Retrieval System	6
1.2.1	Aspects of Information Retrieval Evaluation	6
1.2.2	Precision, Recall, and Their Trade-Offs	7
1.2.3	Ranked Retrieval	9
1.2.4	Standard Test Collections	10
1.3	Exercises	11
2	The Information Retrieval Process	13
2.1	A Bird's Eye View	13
2.1.1	Logical View of Documents	14
2.1.2	Indexing Process	15
2.2	A Closer Look at Text	15
2.2.1	Textual Operations	16
2.2.2	Empirical Laws About Text	18
2.3	Data Structures for Indexing	19
2.3.1	Inverted Indexes	20
2.3.2	Dictionary Compression	21
2.3.3	B and B+ Trees	23
2.3.4	Evaluation of B and B+ Trees	25
2.4	Exercises	25
3	Information Retrieval Models	27
3.1	Similarity and Matching Strategies	27
3.2	Boolean Model	28

3.2.1	Evaluating Boolean Similarity	28
3.2.2	Extensions and Limitations of the Boolean Model	29
3.3	Vector Space Model	30
3.3.1	Evaluating Vector Similarity	30
3.3.2	Weighting Schemes and $tf \times idf$	31
3.3.3	Evaluation of the Vector Space Model	32
3.4	Probabilistic Model	32
3.4.1	Binary Independence Model	33
3.4.2	Bootstrapping Relevance Estimation	34
3.4.3	Iterative Refinement and Relevance Feedback	35
3.4.4	Evaluation of the Probabilistic Model	36
3.5	Exercises	36
4	Classification and Clustering	39
4.1	Addressing Information Overload with Machine Learning	39
4.2	Classification	40
4.2.1	Naive Bayes Classifiers	41
4.2.2	Regression Classifiers	42
4.2.3	Decision Trees	43
4.2.4	Support Vector Machines	44
4.3	Clustering	45
4.3.1	Data Processing	46
4.3.2	Similarity Function Selection	46
4.3.3	Cluster Analysis	48
4.3.4	Cluster Validation	51
4.3.5	Labeling	52
4.4	Application Scenarios for Clustering	53
4.4.1	Search Results Clustering	53
4.4.2	Database Clustering	55
4.5	Exercises	56
5	Natural Language Processing for Search	57
5.1	Challenges of Natural Language Processing	57
5.1.1	Dealing with Ambiguity	58
5.1.2	Leveraging Probability	58
5.2	Modeling Natural Language Tasks with Machine Learning	59
5.2.1	Language Models	59
5.2.2	Hidden Markov Models	60
5.2.3	Conditional Random Fields	60
5.3	Question Answering Systems	61
5.3.1	What Is Question Answering?	61
5.3.2	Question Answering Phases	62
5.3.3	Deep Question Answering	64
5.3.4	Shallow Semantic Structures for Text Representation	66
5.3.5	Answer Reranking	67
5.4	Exercises	68

Part II Information Retrieval for the Web		
6	Search Engines	71
6.1	The Search Challenge	71
6.2	A Brief History of Search Engines	72
6.3	Architecture and Components	74
6.4	Crawling	75
6.4.1	Crawling Process	76
6.4.2	Architecture of Web Crawlers	78
6.4.3	DNS Resolution and URL Filtering	80
6.4.4	Duplicate Elimination	80
6.4.5	Distribution and Parallelization	81
6.4.6	Maintenance of the URL Frontier	82
6.4.7	Crawling Directives	84
6.5	Indexing	85
6.5.1	Distributed Indexing	87
6.5.2	Dynamic Indexing	88
6.5.3	Caching	89
6.6	Exercises	90
7	Link Analysis	91
7.1	The Web Graph	91
7.2	Link-Based Ranking	93
7.3	PageRank	94
7.3.1	Random Surfer Interpretation	96
7.3.2	Managing Dangling Nodes	97
7.3.3	Managing Disconnected Graphs	99
7.3.4	Efficient Computation of the PageRank Vector	100
7.3.5	Use of PageRank in Google	101
7.4	Hypertext-Induced Topic Search (HITS)	101
7.4.1	Building the Query-Induced Neighborhood Graph	102
7.4.2	Computing the Hub and Authority Scores	103
7.4.3	Uniqueness of Hub and Authority Scores	107
7.4.4	Issues in HITS Application	108
7.5	On the Value of Link-Based Analysis	109
7.6	Exercises	110
8	Recommendation and Diversification for the Web	111
8.1	Pruning Information	111
8.2	Recommendation Systems	112
8.2.1	User Profiling	112
8.2.2	Types of Recommender Systems	113
8.2.3	Content-Based Recommendation Techniques	113
8.2.4	Collaborative Filtering Techniques	114
8.3	Result Diversification	116
8.3.1	Scope	116
8.3.2	Diversification Definition	116

8.3.3	Diversity Criteria	117
8.3.4	Balancing Relevance and Diversity	117
8.3.5	Diversification Approaches	118
8.3.6	Multi-domain Diversification	119
8.4	Exercises	120
9	Advertising in Search	121
9.1	Web Monetization	121
9.2	Advertising on the Web	121
9.3	Terminology of Online Advertising	124
9.4	Auctions	125
9.4.1	First-Price Auctions	126
9.4.2	Second-Price Auctions	127
9.5	Pragmatic Details of Auction Implementation	129
9.6	Federated Advertising	130
9.7	Exercises	132
Part III Advanced Aspects of Web Search		
10	Publishing Data on the Web	137
10.1	Options for Publishing Data on the Web	137
10.2	The Deep Web	139
10.3	Web APIs	142
10.4	Microformats	145
10.5	RDFa	148
10.6	Linked Data	152
10.7	Conclusion and Outlook	156
10.8	Exercises	158
11	Meta-search and Multi-domain Search	161
11.1	Introduction and Motivation	161
11.2	Top- <i>k</i> Query Processing over Data Sources	162
11.2.1	OID-Based Problem	163
11.2.2	Attribute-Based Problem	166
11.3	Meta-search	168
11.4	Multi-domain Search	171
11.4.1	Service Registration	171
11.4.2	Processing Multi-domain Queries	173
11.4.3	Exploratory Search	175
11.4.4	Data Visualization	177
11.5	Exercises	178
12	Semantic Search	181
12.1	Understanding Semantic Search	181
12.2	Semantic Model	184
12.3	Resources	188
12.3.1	System Perspective	188

12.3.2	User Perspective	190
12.4	Queries	190
12.4.1	User Perspective	192
12.4.2	System Perspective	192
12.4.3	Query Translation and Presentation	194
12.5	Semantic Matching	195
12.6	Constructing the Semantic Model	198
12.7	Semantic Resources Annotation	202
12.8	Conclusions and Outlook	204
12.9	Exercises	205
13	Multimedia Search	207
13.1	Motivations and Challenges of Multimedia Search	207
13.1.1	Requirements and Applications	207
13.1.2	Challenges	209
13.2	MIR Architecture	211
13.2.1	Content Process	213
13.2.2	Query Process	214
13.3	MIR Metadata	216
13.4	MIR Content Processing	217
13.5	Research Projects and Commercial Systems	218
13.5.1	Research Projects	218
13.5.2	Commercial Systems	220
13.6	Exercises	221
14	Search Process and Interfaces	223
14.1	Search Process	223
14.2	Information Seeking Paradigms	225
14.3	User Interfaces for Search	228
14.3.1	Query Specification	228
14.3.2	Result Presentation	230
14.3.3	Faceted Search	233
14.4	Exercises	234
15	Human Computation and Crowdsearching	235
15.1	Introduction	235
15.1.1	Background	236
15.2	Applications	238
15.2.1	Games with a Purpose	238
15.2.2	Crowdsourcing	240
15.2.3	Human Sensing and Mobilization	242
15.3	The Human Computation Framework	244
15.3.1	Phases of Human Computation	244
15.3.2	Human Performers	246
15.3.3	Examples of Human Computation	246
15.3.4	Dimensions of Human Computation Applications	249

15.4 Research Challenges and Projects	250
15.4.1 The CrowdSearcher Project	250
15.4.2 The CUBRIK Project	252
15.5 Open Issues	256
15.6 Exercises	257
References	259
Index	277