

Contents

1 Introduction	1
1.1 The Growth of the Digital Universe	1
1.2 The Terminology	5
1.3 A Brief History	5
1.3.1 Enabling Technologies and Standards	6
1.3.2 Publicly-Funded Research Initiatives	7
1.3.3 Conferences and Evaluation Campaigns	9
1.3.4 Commercial Products	10
1.4 The Current Research Challenges	13
References	15
2 Within-Language Information Retrieval	17
2.1 Introduction	17
2.2 The Retrieval Problem and Its Consequences	20
2.3 Implementation of a Within-Language Information Retrieval System	22
2.4 Indexing Phase	24
2.4.1 Pre-processing (Step 1)	25
2.4.2 Language Identification (Step 2)	26
2.4.3 Document Formation (Step 3)	27
2.4.4 Segmentation, Tokenisation, Parsing (Step 4)	28
2.4.5 Feature Normalisation (Step 5)	31
2.4.6 Enrichment (Step 6)	35
2.5 Matching Phase	36
2.5.1 'Bag of Words' Paradigm	36
2.5.2 Inverted Index	38
2.5.3 Basic Matching Algorithm	39
2.5.4 Vector Space Model	41
2.5.5 The tf.idf-Cosine Weighting Scheme	43
2.5.6 Relevance Feedback	46

2.5.7 Probabilistic Weighting Schemes	48
2.5.8 Ranking Using Language Models	50
2.5.9 Off-Page Information: Page Rank	51
2.6 Summary and Future Directions	52
2.7 Suggested Reading	53
References	54
3 Cross-Language Information Retrieval	57
3.1 Introduction	57
3.2 Implementation of Cross-Language Information Retrieval	58
3.2.1 Query Translation and Document Translation	58
3.2.2 No Translation	59
3.2.3 Different Types of Translation Resources	60
3.2.4 Term Ambiguity	61
3.3 Translation Approaches for Cross-Language Information Retrieval	62
3.3.1 Machine-Readable Dictionaries	62
3.3.2 Statistical Approaches	65
3.3.3 Pre-translation and Post-translation Query Expansion	69
3.3.4 Machine Translation	71
3.3.5 Combination Approaches	71
3.4 Handling Many Languages	73
3.4.1 CLIR Flows	73
3.4.2 Merging Across Languages	77
3.4.3 Document Translation	79
3.4.4 Indirect Translation	79
3.5 Summary and Future Directions	80
3.6 Suggested Reading	82
References	83
4 Interaction and User Interfaces	85
4.1 Information Seeking and User Interaction	85
4.2 Users' Information Needs and Search Tasks	89
4.3 Users' Language Skills and Cultural Differences	92
4.4 Supporting Multilingual User Interaction	94
4.4.1 Query Formulation and Translation	95
4.4.2 Document Selection and Examination	102
4.4.3 Query Reformulation	108
4.4.4 Browsing and Visualisation	109
4.5 Designing Multilingual Search User Interfaces	114
4.5.1 User-Centred Design	116
4.5.2 Internationalisation and Localisation	118
4.5.3 Case Study: CLIR in Google's Web Search	121
4.6 Summary and Future Directions	123
4.7 Suggested Reading	124
References	125

5 Evaluation for Multilingual Information Retrieval Systems	129
5.1 Introduction	129
5.2 System-Oriented Evaluation	130
5.2.1 The Cranfield Tradition	131
5.2.2 Evaluation Campaigns	132
5.2.3 Building a Test Collection	133
5.2.4 Promoting Research into Multilingual and Multimedia System Development via Evaluation	141
5.2.5 Alternative Methodologies for Test Collection Construction	143
5.2.6 Performance Measures	145
5.2.7 Statistical Significance Testing	154
5.2.8 System Effectiveness and User Satisfaction	155
5.3 User-Oriented Evaluation	156
5.3.1 Experimental Design	157
5.3.2 Evaluating Interactive CLIR Systems at CLEF	159
5.3.3 Alternative Performance Measures	160
5.4 Evaluating Your Own System	161
5.5 Summary and Future Directions	164
5.6 Suggested Reading	165
References	166
6 Applications of Multilingual Information Access	171
6.1 Introduction	171
6.2 Beyond Multilingual Textual Document Retrieval	172
6.2.1 Image Retrieval	173
6.2.2 Speech Retrieval	177
6.2.3 Video Retrieval	180
6.2.4 Question Answering	183
6.3 Multilingual Information Access in Practice	188
6.3.1 Web Search	189
6.3.2 Digital Libraries and Cultural Heritage	191
6.3.3 Medicine and Healthcare	195
6.3.4 Government and Law	196
6.3.5 Business and Commerce	199
6.4 Summing Up	201
References	202
Glossary of Acronyms	209
Index	213