Contents

Dedication v
List of Figures xi
Preface xv
Acknowledgments xix

1. INTRODUCTION 1

2. RETRIEVAL STRATEGIES 9
   2.1 Vector Space Model 11
   2.2 Probabilistic Retrieval Strategies 21
   2.3 Language Models 45
   2.4 Inference Networks 57
   2.5 Extended Boolean Retrieval 67
   2.6 Latent Semantic Indexing 70
   2.7 Neural Networks 74
   2.8 Genetic Algorithms 80
   2.9 Fuzzy Set Retrieval 84
   2.10 Summary 90
   2.11 Exercises 91

3. RETRIEVAL UTILITIES 93
   3.1 Relevance Feedback 94
   3.2 Clustering 105
   3.3 Passage-based Retrieval 113
   3.4 N-grams 115
   3.5 Regression Analysis 119
3.6 Thesauri 122
3.7 Semantic Networks 132
3.8 Parsing 139
3.9 Summary 146
3.10 Exercises 146

4. CROSS-LANGUAGE INFORMATION RETRIEVAL 149
4.1 Introduction 149
4.2 Crossing the Language Barrier 151
4.3 Cross-Language Retrieval Strategies 157
4.4 Cross Language Utilities 170
4.5 Summary 178
4.6 Exercises 179

5. EFFICIENCY 181
5.1 Inverted Index 182
5.2 Query Processing 195
5.3 Signature Files 199
5.4 Duplicate Document Detection 203
5.5 Summary 208
5.6 Exercises 209

6. INTEGRATING STRUCTURED DATA AND TEXT 211
6.1 Review of the Relational Model 215
6.2 A Historical Progression 222
6.3 Information Retrieval as a Relational Application 228
6.4 Semi-Structured Search using a Relational Schema 245
6.5 Multi-dimensional Data Model 250
6.6 Mediators 250
6.7 Summary 253
6.8 Exercises 254

7. PARALLEL INFORMATION RETRIEVAL 257
7.1 Parallel Text Scanning 258
7.2 Parallel Indexing 263
7.3 Clustering and Classification 270
7.4 Large Parallel Systems 271
Contents

7.5 Summary 272
7.6 Exercises 274

8. DISTRIBUTED INFORMATION RETRIEVAL 275
  8.1 A Theoretical Model of Distributed Retrieval 276
  8.2 Web Search 281
  8.3 Result Fusion 284
  8.4 Peer-to-Peer Information Systems 286
  8.5 Other Architectures 289
  8.6 Summary 290
  8.7 Exercises 290

9. SUMMARY AND FUTURE DIRECTIONS 291

References 299

Index 331