Text Information Retrieval Systems

Second Edition

Charles T. Meadow
Bert R. Boyce
Donald H. Kraft

Contents

Preface xv

1 Introduction

1.1 What Is Information? 1
1.2 What Is Information Retrieval? 2
1.3 How Does Information Retrieval Work? 4
1.3.1 The User Sequence 6
1.3.2 The Database Producer Sequence 10
1.3.3 Query Processing 13
1.3.4 Why the Process Is Not Perfect 13
1.4 Who Uses Information Retrieval? 15
1.4.1 Information Specialists 16
1.4.2 End Users 16
1.4.3 Novices 17
1.5 What Are the Problems in IRS Design and Use? 17
1.5.1 Design 18
1.5.2 Understanding User Behavior 19
1.6 A Brief History of Information Retrieval 19
1.6.1 Traditional Information Retrieval Methods 20
1.6.2 Pre-computer IR Systems 22
1.6.3 Special Purpose Computer Systems 24
1.6.4 General Purpose Computer Systems 28
1.6.5 Online Database Services 28
1.6.6 Modern Experiments (SMART and Other Systems) 31
1.6.7 The World Wide Web 32

2
Data, Information, and Knowledge

2.1 Introduction 34
2.2 Definitions 35
2.2.1 Data 35
2.2.2 Information 35
2.2.3 News 37
2.2.4 Knowledge 38
2.2.5 Intelligence 39
2.2.6 Meaning 39
2.2.7 Wisdom 40
2.3 Metadata 41
2.4 Knowledge Base 43
2.5 *Credence, Justified Belief, and Point of View* 46
2.6 Summary 48

3
Representation of Information

3.1 Information to Be Represented 49
3.2 Types of Representation 53
3.2.1 Natural Language 54
3.2.2 Restricted Natural Language 55
3.2.3 Artificial Language 56
3.2.4 Codes, Measures, and Descriptors 57
3.2.5 Mathematical Models of Text 58
3.3 Characteristics of Information Representations 60
3.3.1 Discriminating Power 60
3.3.2 Identification of Similarity 61
3.3.3 Descriptiveness 61
3.3.4 Ambiguity 62
3.3.5 Conciseness 62
3.4 Relationships among Entities and Attribute Values 62
3.4.1 Hierarchical Codes 63
3.4.2 Measurements 64
3.4.3 Nominal Descriptors 64
3.4.4 Inflected Language 65
3.4.5 Full Text 66
3.4.6 Explicit Pointers and Links 66
3.5 Summary 67

4

Attribute Content and Values

4.1 Types of Attribute Symbols 68
4.1.1 Numbers 69
4.1.2 Character Strings—Names 69
4.1.3 Other Character Strings 70
4.2 Class Relationships 71
4.2.1 Hierarchical Classification 71
4.2.2 Network Relationships 73
4.2.3 Class Membership—Binary, Probabilistic, or Fuzzy 73
4.3 Transformations of Values 77
4.3.1 Transformation of Words by Stemming 77
4.3.2 Sound-Based Transformation of Words 81
4.3.3 Transformation of Words by Meaning 83
4.3.4 Transformation of Graphics 85
4.3.5 Transformation of Sound 87
4.4 Uniqueness of Values 89
4.5 Ambiguity of Attribute Values 91
4.6 Indexing of Text 93
4.7 Control of Vocabulary 95
4.7.1 Elements of Control 95
4.7.2 Relationships among Terms 96
4.8 The Importance of Point of View 97
4.9 Summary 98

5

Models of Virtual Data Structure

5.1 The Concept of Models of Data 99
5.2 Basic Data Elements and Structures 102
5.2.1 Scalar Variables and Constants 103
5.2.2 Vector Variables 103
5.2.3 Structures 103
5.2.4 Arrays 104
5.2.5 Tuples 104
5.2.6 Relations 105
5.2.7 Text 106
5.3 The Common Structural Models 108
5.3.1 The Linear Sequential Model 108
5.3.2 The Relational Model 109
5.3.3 The Hierarchical Model 111

5.3.4 The Network Model 113
5.4 Applications of the Basic Models 114
5.4.1 Hypertext 114
5.4.2 Spreadsheet Files 116
5.5 The Entity–Relationship Model 117
5.6 Summary 121

6

The Physical Structure of Data

6.1 Introduction to Physical Structures 122
6.2 Record Structures and Their Effects 123
6.2.1 Basic Structures 123
6.2.2 Space–Time and Transaction Rate 125
6.3 Basic Concepts of File Structure 127
6.3.1 The Order of Records 127
6.3.2 Finding Records 128
6.4 Organizational Methods 128
6.4.1 Sequential Files 129
6.4.2 Index-File Structures 131
6.4.3 Lists 132
6.4.4 Trees 139
6.4.5 Direct-Access Structures 140
6.5 Parsing of Data Elements 143
6.5.1 Phrase Parsing 144
6.5.2 Word Parsing 144
6.5.3 Word and Phrase Parsing 145
6.6 Combination Structures 145
6.6.1 Nested Indexes 145
6.6.2 Direct Structure with Chains 148
6.6.3 Indexed Sequential Access Method 149
6.7 Summary 150

7

Querying the Information Retrieval System

7.1 Introduction 152
7.2 Language Types 153
7.3 Query Logic 155
7.3.1 Sets and Subsets 156
7.3.2 Relational Statements 156

7.3.3 Boolean Query Logic 157
7.3.4 Ranked and Fuzzy Sets 160
7.3.5 Similarity Measures 162
7.4 Functions Performed 163
7.4.1 Connect to a Remote IRS 163
7.4.2 Select Database 164
7.4.3 Search the Inverted File or Thesaurus 164
7.4.4 Create a Subset of the Database 165
7.4.5 Search for Strings 165
7.4.6 Analyze a Set 168
7.4.7 Sort, Display, and Format Records 169
7.4.8 Handling the Unstructured Record 170
7.4.9 Download 170
7.4.10 Order Documents 171
7.4.11 Save, Recall, and Edit Searches 172
7.4.12 Current Awareness Search 172
7.4.13 Cost Summary 173
7.4.14 Terminate a Session 174
7.5 The Basis for Charging for Searches 174

8

Interpretation and Execution of Query Statements

8.1 Problems of Query Language Interpretation 176
8.1.1 Parsing Command Language 177
8.1.2 Parsing Natural Language 180
8.1.3 Processing Menu Choices 182
8.2 Executing Retrieval Commands 183
8.2.1 Database Selection 184
8.2.2 Inverted File Search 184
8.2.3 Set or Subset Creation 184
8.2.4 Truncation and Universal Characters 185
8.2.5 Left-Hand Truncation 187
8.3 Executing Record Analysis and Presentation Commands 191
8.3.1 Set Analysis Functions 191
8.3.2 Display, Format, and Sort 192
8.3.3 Offline Printing 195
8.4 Executing Other Commands 195
8.4.1 Ordering 195
8.4.2 Save, Recall, and Edit Searches 196
8.4.3 Current Awareness 196
8.4.4 Cost Summation and Billing 197

8.4.5 Terminating a Session 198
8.5 Feedback to Users and Error Messages 199
8.5.1 Response to Command Errors 199
8.5.2 Set-Size Indication 200
8.5.3 Record Display 200
8.5.4 Set Analysis 200
8.5.5 Cost 200
8.5.6 Help 200

9 Text Searching

9.1 The Special Problems of Text Searching 203
9.2 Some Characteristics of Text and Their Applications 204
9.2.1 Components of Text 204
9.2.2 Significant Words—Indexing 205
9.2.3 Significant Sentences—Abstracting 207
9.2.4 Measures of Complete Texts 210
9.3 Command Language for Text Searching 211
9.3.1 Set Membership Statements 211
9.3.2 Word or String Occurrence Statements 211
9.3.3 Proximity Statements 212
9.4 Term Weighting 215
9.4.1 Indexing with Weights 216
9.4.2 Automated Assignment of Weights 217
9.4.3 Improving Weights 217
9.5 Word Association Techniques 218
9.5.1 Dictionaries and Thesauri 218
9.5.2 Minithesauri 219
9.5.3 Word Co-occurrence Statistics 220
9.5.4 Stemming and Conflation 221
9.6 Text or Record Association Techniques 221
9.6.1 Similarity Measures 222
9.6.2 Clustering 224
9.6.3 Signature Matching 228
9.6.4 Discriminant Methods 230
9.7 Other Processes with Words of a Text 232
9.7.1 Stop Words 232
9.7.2 Replacement of Words with Roots or Associated Words 233
9.7.3 Varying Significance as a Function of Frequency 233

9.7.4 Comments on the Computation of the Strength of Document Association 234

10 System-Computed Relevance and Ranking

10.1 The Retrieval Status Value (rsv) 238
10.2 Ranking 239
10.3 Methods of Evaluating the rsv 239
10.3.1 The Vector Space Model 239
10.3.2 The Probabilistic Model 240
10.3.3 The Extended Boolean Model 242

11 Search Feedback and Iteration

11.1 Basic Concepts of Feedback and Iteration 246
11.2 Command Sequences 247
11.3 Information Available as Feedback 249
11.3.1 File or Database Selection 249
11.3.2 Term Search or Browsing 249
11.3.3 Record Search or Set Formation 251
11.3.4 Record Display and Browsing 252
11.3.5 Record Acquisition 252
11.3.6 Requests for Information about the Retrieval System 254
11.3.7 Establishing Communication Parameters 254
11.3.8 Trends over Sequences and Cycles 255
11.4 Adjustments in the Search 255
11.4.1 Improve Term Selection 256
11.4.2 Improve Set Formation Logic 256
11.4.3 Improve Final Set Size 257
11.4.4 Improve Precision, Recall, or Total Utility 257
11.5 Feedback from User to System 257

12 Multidatabase Searching and Mapping

12.1 Basic Concepts 261
12.2 Multidatabase Search 262
12.2.1 The Nature of Duplicate Records 262

12.2.2 Detection of Duplicates 264
12.2.3 Scanning Multiple Databases 266
12.3 Mapping 268
12.4 Value of Mapping 270

13

Search Strategy

13.1 The Nature of Searching Reconsidered 272
13.1.1 Known Item Search 273
13.1.2 Specific Information Search 273
13.1.3 General Information Search 274
13.1.4 Exploration of the Database 274
13.2 The Nature of Search Strategy 275
13.2.1 Search Objective 275
13.2.2 General Plan of Operation 276
13.2.3 The Essential Information Elements of a Search 276
13.2.4 Specific Plan of Operation 277
13.3 Types of Strategies 278
13.3.1 Categorizing by Objective 278
13.3.2 Categorizing by Plan of Operation 279
13.4 Tactics 281
13.4.1 Monitoring Tactics 282
13.4.2 File Structure Tactics 282
13.4.3 Search Formulation Tactics 282
13.4.4 Term Tactics 282
13.5 Summary 282

14

The Information Retrieval System Interface

14.1 General Model of Message Flow 283
14.2 Sources of Ambiguity 286
14.3 The Role of a Search Intermediary 287
14.3.1 Establishing the Information Need 288
14.3.2 Development of a Search Strategy 289
14.3.3 Translation of the Need Statement into a Query 289
14.3.4 Interpretation and Evaluation of Output 289
14.3.5 Search Iteration within the Strategic Plan 289
14.3.6 Change of Strategy When Necessary 290
14.3.7 Help in Using an IRS 290

14.4 Automated Search Mediation 290
14.4.1 Early Development 291
14.4.2 Fully Automatic Intermediary Functions 291
14.4.3 Interactive Intermediary Functions 292
14.5 The User Interface as a Component of All Systems 295

15

A Sampling of Information Retrieval Systems

15.1 Introduction 296
15.2 Dialog 296
15.2.1 A Command Language Using Boolean Logic 297
15.2.2 Target 299
15.2.3 DIALOGweb–A Web Adaptation 299
15.3 AltaVista 302
15.3.1 Default Query Entry Form 303
15.3.2 Advanced Search Form 305
15.4 Northern Light Technology 305
15.5 The *Canadian Encyclopedia* and X-Portal™ 306
15.5.1 The Encyclopedia 307
15.5.2 X-Portal™ 308
15.6 Summary 309

16

Measurement and Evaluation

16.1 Basics of Measurement 310
16.1.1 The Data Manager 311
16.1.2 The Query Manager 312
16.1.3 The Query Composition Process 312
16.1.4 Deriving the Information Need 313
16.1.5 The Database 313
16.1.6 Users 314
16.2 Relevance, Value, and Utility 314
16.2.1 Relevance as Relatedness 315
16.2.2 Aspects of Value 315
16.2.3 Relevance as Utility 316
16.2.4 Retaining Two Separate Measures 317
16.2.5 The Relevance Measurement Scale 319
16.2.6 Taking the Measurements 319
16.2.7 Questions about Relevance as a Measure 321

16.3 Measures Based on Relevance 321
16.3.1 Precision (*P*) 322
16.3.2 Recall (*R*) 323
16.3.3 Relationship of *Recall* and *Precision* 323
16.3.4 Overall Effectiveness Measures Based on *R* and *P* 325
16.4 Measures of Process 328
16.4.1 Query Translation 328
16.4.2 Errors in a Query Statement 328
16.4.3 Average Time per Command or per User Decision 329
16.4.4 Elapsed Time of a Search 329
16.4.5 Number of Commands or Steps in a Search 329
16.4.6 Cost of a Search 329
16.4.7 Size of Final Set Formed 330
16.4.8 Number of Records Reviewed by the User 330
16.4.9 Patterns of Language Use 330
16.5 Measures of Outcome 333
16.5.1 Precision 335
16.5.2 Recall 335
16.5.3 Efficiency 335
16.5.4 Overall User Evaluation 335
16.6 Measures of Environment 336
16.6.1 Database Record Selection 336
16.6.2 Record Content 336
16.6.3 Measures of Users 337
16.7 Conclusion 337

Bibliography 338

Recommended Reading 349

Index 351