INTERNATIONAL SERIES IN OPERATIONS RESEARCH AND MANAGEMENT SCIENCE

Data Engineering

Mining, Information and Intelligence

Yupo Chan
John Talburt
Terry M. Talley
Editors

Springer

Table of Contents

1 Introduction **1**
1.1 Common Problem 1
1.2 Data Integration and Data Management 3
1.2.1 Information Quality Overview 3
1.2.2 Customer Data Integration 4
1.2.3 Data Management 8
1.2.4 Practical Problems to Data Integration and Management 9
1.3 Analytics 10
1.3.1 Model Development 10
1.3.2 Current Modeling and Optimization Techniques 11
1.3.3 Specific Algorithms and Techniques for Improvement 12
1.3.4 Incremental or Evolutionary Updates 13
1.3.5 Visualization 15
1.4 Conclusion 15
1.5 References 16

2 A Declarative Approach to Entity Resolution **17**
2.1 Introduction 17
2.2 Background 18
2.2.1 Entity Resolution Definition 18
2.2.2 Entity Resolution Defense 18
2.2.3 Entity Resolution Terminology 19
2.2.4 Declarative Languages 20
2.3 The Declarative Taxonomy: The Nouns 20
2.3.1 Attributes 21
2.3.2 References 21
2.3.3 Paths and Match Functions 22
2.3.4 Entities 24
2.3.5 Super Groups 25
2.3.6 Matching Graphs 26
2.4 A Declarative Taxonomy: The Adjectives 27
2.4.1 Attribute Adjectives 27
2.4.2 Reference Adjectives 29
2.5 The Declarative Taxonomy: The Verbs 29
2.5.1 Attribute Verbs 29
2.5.2 Reference Verbs 30
2.5.3 Entity Verbs 32

2.6 A Declarative Representation 33
2.6.1 The XML Schema 34
2.6.2 A Representation for the Operations 36
2.7 Conclusion 37
2.8 Exercises 37
2.9 References 37

3 Transitive Closure of Data Records: Application and Computation 39
3.1 Introduction 39
3.1.1 Motivation 40
3.1.2 Literature Review 42
3.2 Problem Definition 43
3.3 Sequential Algorithms 45
3.3.1 A Breadth First Search Based Algorithm 45
3.3.2 A Sorting and Disjoint Set Based Algorithm 47
3.3.3 Experiment 51
3.4 Parallel and Distributed Algorithms 53
3.4.1 An Overview of a Parallel and Distributed Scheme 53
3.4.2 Generate Matching Pairs 55
3.4.3 Conversion Process 55
3.4.4 Closure Process 56
3.4.5 A MPI Based Parallel and Distributed Algorithm 62
3.4.6 Experiment 64
3.5 Conclusion 70
3.6 Exercises 71
3.7 Acknowledgments 73
3.8 References 74

4 Semantic Data Matching: Principles and Performance 77
4.1 Introduction 77
4.2 Problem Statement: Data Matching for Customer Data Integration 78
4.3 Semantic Data Matching 78
4.3.1 Background on Latent Semantic Analysis 78
4.3.2 Analysis 80
4.4 Effect of Shared Terms 81
4.4.1 Fundamental Limitations on Data Matching 81
4.4.2 Experiments 82
4.5 Results 83
4.6 Conclusion 87
4.7 Exercises 89
4.8 Acknowledgments 89
4.9 References 89

5 Application of the Near Miss Strategy and Edit Distance to Handle Dirty Data 91
5.1 Introduction 91
5.2 Background 92

5.2.1 Techniques used for General Spelling Error Correction 93
5.2.2 Domain-Specific Correction 95
5.3 Individual Name Spelling Correction Algorithm: the Personal Name Recognition Strategy (PNRS) 96
5.3.1 Experiment Results 98
5.4 Conclusion 99
5.5 Exercises 99
5.6 References 100

6 A Parallel General-Purpose Synthetic Data Generator 103
6.1 Introduction 103
6.2 SDDL 104
6.2.1 Min/Max Constraints 105
6.2.2 Distribution Constraints 106
6.2.3 Formula Constraints 106
6.2.4 Iterations 106
6.2.5 Query Pools 108
6.3 Pools 108
6.4 Parallel Data Generation 110
6.4.1 Generation Algorithm 1 111
6.4.2 Generation Algorithm 2 112
6.5 Performance and Applications 113
6.6 Conclusion and Future Directions 114
6.7 Exercises 116
6.8 References 117

7 A Grid Operating Environment for CDI 119
7.1 Introduction 119
7.2 Grid-Based Service Deployment 120
7.2.1 Evolution of the Acxiom Grid (A Case Study) 120
7.2.2 Services Grid 122
7.2.3 Grid Management 124
7.3 Grid-Based Batch Processing 127
7.3.1 Workflow Grid 127
7.3.2 I/O Constraints 133
7.3.3 Data Grid 135
7.3.4 Database Grid 137
7.3.5 Data Management 138
7.4 Conclusion 140
7.5 Exercises 141

8 Parallel File Systems 143
8.1 Introduction 143
8.2 Commercial Data and Access Patterns 144
8.2.1 Large File Access Patterns 145
8.2.2 File System Interfaces 146

8.3 Basics of Parallel File Systems 147
8.3.1 Common Storage System Hardware 148
8.4 Design Challenges 149
8.4.1 Performance 150
8.4.2 Consistency Semantics 150
8.4.3 Fault Tolerance 151
8.4.4 Interoperability 152
8.4.5 Management Tools 153
8.4.6 Traditional Design Challenges 154
8.5 Case Studies 154
8.5.1 Multi-Path File System (MPFS) 154
8.5.2 Parallel Virtual File System (PVFS) 157
8.5.3 The Google File System (GFS) 160
8.5.4 pNFS 163
8.6 Conclusion 167
8.7 Exercises 167
8.8 References 168

9 Performance Modeling of Enterprise Grids 169
9.1 Introduction and Background 169
9.1.1 Performance Modeling 169
9.1.2 Capacity Planning Tools and Methodology 171
9.2 Measurement Collection and Preliminary Analysis 173
9.3 Workload Characterization 174
9.3.1 K-means Clustering 176
9.3.2 Hierarchical Workload Characterization 181
9.3.3 Other Issues in Workload Characterization 182
9.4 Baseline System Models and Tool Construction 184
9.4.1 Analytic Models 184
9.4.2 Simulation Tools for Enterprise Grid Systems 191
9.5 Enterprise Grid Capacity Planning Case Study 192
9.5.1 Data Collection and Preliminary Analysis 194
9.5.2 Workload Characterization 194
9.5.3 Development and Validation of the Baseline Model 195
9.5.4 Model Predictions 196
9.6 Summary 199
9.7 Exercises 199
9.8 References 200

10 Delay Characteristics of Packet Switched Networks 203
10.1 Introduction 203
10.2 High-Speed Packet Switching Systems 204
10.2.1 Packet Switched General Organization 204
10.2.2 Switching Fabric Structures for Packet Switches 205
10.2.3 Queuing Schemes for Packet Switches 206
10.3 Technical Background 207
10.3.1 Packet Scheduling in Packet Switches 207
10.3.2 Introduction to Network Calculus 208

10.4 Delay Characteristics of Output Queuing Switches 210
10.4.1 Output Queuing Switch System 210
10.4.2 OQ Switch Modeling and Analysis 211
10.4.3 Output Queuing Emulation for Delay Guarantee 212
10.5 Delay Characteristics of Buffered Crossbar Switches 212
10.5.1 Buffered Crossbar Switch System 212
10.5.2 Modeling Traffic Control in Buffered Crossbar Switches 214
10.5.3 Delay Analysis for Buffered Crossbar Switches 215
10.5.4 Numerical Examples 216
10.6 Delay Comparison of Output Queuing to Buffered Crossbar 217
10.6.1 Maximum Packet Delay Comparison 217
10.6.2 Bandwidth Allocation for Delay Performance Guarantees 218
10.6.3 Numerical Examples 219
10.7 Summary 221
10.8 Exercises 222
10.9 References 222

11 Knowledge Discovery in Textual Databases: A Concept-Association Mining Approach 225
11.1 Introduction 225
11.1.1 Graph Representation 228
11.2 Method 228
11.2.1 Concept Based Association Rule Mining Approach 228
11.2.2 Concept Extraction 229
11.2.3 Mining Concept Associations 231
11.2.4 Generating a Directed Graph of Concept Associations 231
11.3 Experiments and Results 233
11.3.1 Isolated words vs. multi-word concepts 233
11.3.2 New Metrics vs. the Traditional Support & Confidence 235
11.4 Conclusions 240
11.5 Examples 241
11.6 Exercises 242
11.7 References 242

12 Mining E-Documents to Uncover Structures 245
12.1 Introduction 245
12.2 Related Research 246
12.3 Discovery of the Physical Structure 247
12.3.1 Paragraph 247
12.3.2 Heading 248
12.3.3 Table 252
12.3.4 Image 253
12.3.5 Capturing the physical structure of an e-document. 254
12.4 Discovery of the Explicit Terms Using Ontology 263
12.4.1 The Stemmer 264
12.4.2 The Ontology 264
12.4.3 Discovery Process 266

12.5 Discovery of the Logical Structure 268
12.5.1 Segmentation 268
12.5.2 Segments' Relationships 270
12.6 Empirical Results 272
12.7 Conclusions 274
12.8 Exercises 274
12.9 Acknowledgments 276
12.10 References 276

13 Designing a Flexible Framework for a Table Abstraction 279
13.1 Introduction 279
13.2 Analysis of the Table ADT 281
13.3 Formal Design Contracts 283
13.4 Layered Architecture 285
13.5 Client Layer 286
13.5.1 Abstract Predicates for Keys and Records 287
13.5.2 Keys and the Comparable Interface 287
13.5.3 Records and the Keyed Interface 288
13.5.4 Interactions among the Layers 289
13.6 Access Layer 289
13.6.1 Abstract Predicates for Tables 289
13.6.2 Table Interface 289
13.6.3 Interactions among the Layers 291
13.7 Storage Layer 292
13.7.1 Abstract Predicate for Storable Records 292
13.7.2 Bridge Pattern 292
13.7.3 Proxy Pattern 293
13.7.4 RecordStore Interface 294
13.7.5 RecordSlot Interface 295
13.7.6 Interactions among the Layers 297
13.8 Externalization Module 297
13.9 Iterators 299
13.9.1 Table Iterator Methods 300
13.9.2 Input Iterators 301
13.9.3 Filtering Iterators 302
13.9.4 Query Iterator Methods 303
13.10 Evolving Frameworks 305
13.10.1 Three Examples 305
13.10.2 Whitebox Frameworks 306
13.10.3 Component Library 306
13.10.4 Hot Spots 307
13.10.5 Pluggable Objects 308
13.11 Discussion 308
13.12 Conclusion 310
13.13 Exercises 310
13.14 Acknowledgements 312
13.15 References 312

14 Information Quality Framework for Verifiable Intelligence Products ...315
14.1 Introduction ... 315
14.2 Background ... 317
14.2.1 Production Process of Intelligence Products ... 317
14.2.2 Current IQ Practices in the IC ... 319
14.2.3 Relevant Concepts and Methods of IQ Management ... 321
14.3 IQ Challenges within the IC ... 323
14.3.1 IQ Issues in Intelligence Collection and Analysis ... 323
14.3.2 Other IQ Problems ... 324
14.3.3 IQ Dimensions Related to the IC ... 325
14.4 Towards a Proposed Solution ... 326
14.4.1 IQ Metrics for Intelligence Products ... 327
14.4.2 Verifiability of Intelligence Products ... 328
14.4.3 Objectives and Plan ... 329
14.5 Conclusion ... 331
14.6 Exercises ... 331
14.7 References ... 331

15 Interactive Visualization of Large High-Dimensional Datasets ... 335
15.1 Introduction ... 335
15.1.1 Related work ... 335
15.1.2 General requirements for a data visualization system ... 336
15.2 Data Visualization Process ... 337
15.2.1 Data Rendering Stage ... 338
15.2.2 Backward Transformation Stage ... 341
15.2.3 Knowledge Extraction Stage ... 342
15.3 Interactive Visualization Model ... 343
15.4 Utilizing Summary Icons ... 344
15.5 A Case Study ... 346
15.6 Conclusion ... 350
15.7 Exercises ... 350
15.8 Acknowledgements ... 350
15.9 References ... 350

16 Image Watermarking Based on Pyramid Decomposition with CH Transform ... 353
16.1. Introduction ... 353
16.2. Algorithm for multi-layer image watermarking ... 354
16.2.1. Resistant watermarking ... 354
16.2.2. Resistant watermark detection ... 364
16.2.3. Fragile watermarking ... 369
16.3. Data hiding ... 370
16.4. Evaluation of the watermarking efficiency ... 371
16.5. Experimental results ... 372
16.6. Application areas ... 379
16.6.1. Resistant watermarks ... 379
16.6.2. Fragile watermarks ... 380
16.6.3. Data hiding ... 380

16.7. Conclusion 380
16.8 Exercises 381
16.9 Acknowledgment 386
16.10 References 386

17 Immersive Visualization of Cellular Structures 389
17.1 Introduction 389
17.2 Light Microscopic Cellular Images and Focus: Basics 390
17.3 Flat-Field Correction 392
17.4 Separation of Transparent Layers using Focus 393
17.5 3D Visualization of Cellular Structures 396
17.5.1 Volume Rendering 396
17.5.2 Immersive Visualization: CAVE Environment 398
17.6 Conclusions 401
17.7 Exercises 401
17.8 References 401

18 Visualization and Ontology of Geospatial Intelligence 403
18.1 Introduction 403
18.1.1 Premises 403
18.1.2 Research Agenda 404
18.2 Semantic Information Representation and Extrac on 405
18.3 Markov Random Field 406
18.3.1 Spatial or Contextual Pattern Recognition 407
18.3.2 Image Classification using *k*-medoid Method 407
18.3.3 Random Field and Spatial Time Series 410
18.3.4 First Persian-Gulf-War Example 412
18.4 Context-driven Visualization 414
18.4.1 Relevant Methodologies 414
18.4.2 Visual Perception and Tracking 415
18.4.3 Visualization 417
18.5 Intelligent Information Fusion 419
18.5.1 Semantic Information Extraction 419
18.5.2 Intelligent Contextual Inference 420
18.5.3 Context-driven Ontology 420
18.6 Metrics for Knowledge Extraction and Discovery 421
18.7 Conclusions and Recommendations 422
18.7.1 Contributions 422
18.7.2 Looking Ahead 423
18.8 Exercises 424
18.9 Acknowledgements 427
18.10 References 428

19 Looking Ahead 431
19.1 Introduction 431
19.2 Data Integration and Information Quality 432
19.3 Grid Computing 434

19.4 Data Mining 435
19.5 Visualization 437
19.6 References 438

Index 441