Contents

Preface xiii

List of Contributors xvii

1  Data mining meets grid computing: Time to dance? 1
   Alberto Sánchez, Jesús Montes, Werner Dubitzky, Julio J. Valdés, María S. Pérez and Pedro de Miguel
   1.1 Introduction 2
   1.2 Data mining 3
       1.2.1 Complex data mining problems 3
       1.2.2 Data mining challenges 4
   1.3 Grid computing 6
       1.3.1 Grid computing challenges 9
   1.4 Data mining grid – mining grid data 9
       1.4.1 Data mining grid: a grid facilitating large-scale data mining 9
       1.4.2 Mining grid data: analyzing grid systems with data mining techniques 11
   1.5 Conclusions 12
   1.6 Summary of Chapters in this Volume 13

2  Data analysis services in the knowledge grid 17
   Eugenio Cesario, Antonio Congiusta, Domenico Talia and Paolo Trunfio
   2.1 Introduction 17
   2.2 Approach 18
   2.3 Knowledge Grid services 20
       2.3.1 The Knowledge Grid architecture 21
       2.3.2 Implementation 24
   2.4 Data analysis services 29
   2.5 Design of Knowledge Grid applications 31
       2.5.1 The VEGA visual language 31
       2.5.2 UML application modelling 32
       2.5.3 Applications and experiments 33
   2.6 Conclusions 34
3 GridMiner: An advanced support for e-science analytics

Peter Brezany, Ivan Janciak and A. Min Tjoa

3.1 Introduction                          37
3.2 Rationale behind the design and development of GridMiner  39
3.3 Use Case                              40
3.4 Knowledge discovery process and its support by the GridMiner  41
    3.4.1 Phases of knowledge discovery  42
    3.4.2 Workflow management            45
    3.4.3 Data management                46
    3.4.4 Data mining services and OLAP  47
    3.4.5 Security                       49
3.5 Graphical user interface              50
3.6 Future developments                   52
    3.6.1 High-level data mining model   52
    3.6.2 Data mining query language     52
    3.6.3 Distributed mining of data streams 52
3.7 Conclusions                          53

4 ADaM services: Scientific data mining in the service-oriented architecture paradigm

Rahul Ramachandran, Sara Graves, John Rushing, Ken Keyzer, Manil Maskey, Hong Lin and Helen Conover

4.1 Introduction                          58
4.2 ADaM system overview                 58
4.3 ADaM toolkit overview                60
4.4 Mining in a service-oriented architecture 61
4.5 Mining web services                  62
    4.5.1 Implementation architecture   63
    4.5.2 Workflow example              64
    4.5.3 Implementation issues         64
4.6 Mining grid services                  66
    4.6.1 Architecture components       67
    4.6.2 Workflow example              68
4.7 Summary                               69

5 Mining for misconfigured machines in grid systems

Noam Palatin, Arie Leizarowitz, Assaf Schuster and Ran Wolff

5.1 Introduction                          71
5.2 Preliminaries and related work        73
    5.2.1 System misconfiguration detection 73
    5.2.2 Outlier detection                74
5.3 Acquiring, pre-processing and storing data 75
    5.3.1 Data sources and acquisition    75
    5.3.2 Pre-processing                  75
    5.3.3 Data organization               76
6 FAEHIM: Federated Analysis Environment for Heterogeneous Intelligent Mining

Ali Shaikh Ali and Omer F. Rana

6.1 Introduction 91
6.2 Requirements of a distributed knowledge discovery framework 93
  6.2.1 Category 1: knowledge discovery specific requirements 93
  6.2.2 Category 2: distributed framework specific requirements 94
6.3 Workflow-based knowledge discovery 94
6.4 Data mining toolkit 95
6.5 Data mining service framework 96
6.6 Distributed data mining services 99
6.7 Data manipulation tools 100
6.8 Availability 101
6.9 Empirical experiments 101
  6.9.1 Evaluating the framework accuracy 102
  6.9.2 Evaluating the running time of the framework 103
6.10 Conclusions 104

7 Scalable and privacy preserving distributed data analysis over a service-oriented platform

William K. Cheung

7.1 Introduction 105
7.2 A service-oriented solution 106
7.3 Background 107
  7.3.1 Types of distributed data analysis 107
  7.3.2 A brief review of distributed data analysis 108
  7.3.3 Data mining services and data analysis management systems 108
7.4 Model-based scalable, privacy preserving, distributed data analysis 109
  7.4.1 Hierarchical local data abstractions 109
  7.4.2 Learning global models from local abstractions 110
7.5 Modelling distributed data mining and workflow processes 111
  7.5.1 DDM processes in BPEL4WS 111
  7.5.2 Implementation details 112
7.6 Lessons learned 112
7.6.1 Performance of running distributed data analysis on BPEL 112
7.6.2 Issues specific to service-oriented distributed data analysis 113
7.6.3 Compatibility of Web services development tools 114

7.7 Further research directions 114
7.7.1 Optimizing BPEL4WS process execution 114
7.7.2 Improved support of data analysis process management 115
7.7.3 Improved support of data privacy preservation 115

7.8 Conclusions 116

8 Building and using analytical workflows in Discovery Net 119
Moustafa Ghanem, Vasa Curcin, Patrick Wendel and Yike Guo 119

8.1 Introduction 119
8.1.1 Workflows on the grid 120

8.2 Discovery Net system 121
8.2.1 System overview 121
8.2.2 Workflow representation in DPML 122
8.2.3 Multiple data models 123
8.2.4 Workflow-based services 123
8.2.5 Multiple execution models 123
8.2.6 Data flow pull model 124
8.2.7 Streaming and batch transfer of data elements 124
8.2.8 Control flow push model 125
8.2.9 Embedding 125

8.3 Architecture for Discovery Net 126
8.3.1 Motivation for a new server architecture 126
8.3.2 Management of hosting environments 127
8.3.3 Activity management 127
8.3.4 Collaborative workflow platform 127
8.3.5 Architecture overview 127
8.3.6 Activity service definition layer 129
8.3.7 Activity services bus 130
8.3.8 Collaboration and execution services 130
8.3.9 Workflow Services Bus 130
8.3.10 Prototyping and production clients 130

8.4 Data management 131

8.5 Example of a workflow study 133
8.5.1 ADR studies 133
8.5.2 Analysis overview 133
8.5.3 Service for transforming event data into patient annotations 134
8.5.4 Service for defining exclusions 134
8.5.5 Service for defining exposures 135
8.5.6 Service for building the classification model 135
8.5.7 Validation service 135
8.5.8 Summary 136

8.6 Future directions 136
9 Building workflows that traverse the bioinformatics data landscape

Robert Stevens, Paul Fisher, Jun Zhao, Carole Goble and Andy Brass

9.1 Introduction
9.2 The bioinformatics data landscape
9.3 The bioinformatics experiment landscape
9.4 Taverna for bioinformatics experiments
  9.4.1 Three-tiered enactment in Taverna
  9.4.2 The open-typing data models
9.5 Building workflows in Taverna
  9.5.1 Designing a SCUFL workflow
9.6 Workflow case study
  9.6.1 The bioinformatics task
  9.6.2 Current approaches and issues
  9.6.3 Constructing workflows
  9.6.4 Candidate genes involved in trypanosomiasis resistance
  9.6.5 Workflows and the systematic approach
9.7 Discussion

10 Specification of distributed data mining workflows with DataMiningGrid

Dennis Wegener and Michael May

10.1 Introduction
10.2 DataMiningGrid environment
  10.2.1 General architecture
  10.2.2 Grid environment
  10.2.3 Scalability
  10.2.4 Workflow environment
10.3 Operations for workflow construction
  10.3.1 Chaining
  10.3.2 Looping
  10.3.3 Branching
  10.3.4 Shipping algorithms
  10.3.5 Shipping data
  10.3.6 Parameter variation
  10.3.7 Parallelization
10.4 Extensibility
10.5 Case studies
  10.5.1 Evaluation criteria and experimental methodology
  10.5.2 Partitioning data
  10.5.3 Classifier comparison scenario
  10.5.4 Parameter optimization
10.6 Discussion and related work
10.7 Open issues
10.8 Conclusions
11 Anteater: Service-oriented data mining 179
Renato A. Ferreira, Dorgival O. Guedes and Wagner Meira Jr.

11.1 Introduction 179
11.2 The architecture 181
11.3 Runtime framework 183
   11.3.1 Labelled stream 185
   11.3.2 Global persistent storage 185
   11.3.3 Termination detection 186
   11.3.4 Application of the model 187
11.4 Parallel algorithms for data mining 189
   11.4.1 Decision trees 189
   11.4.2 Clustering 193
11.5 Visual metaphors 195
11.6 Case studies 196
11.7 Future developments 197
11.8 Conclusions and future work 198

12 DMGA: A generic brokering-based Data Mining Grid Architecture 201
Alberto Sánchez, María S. Pérez, Pierre Gueant, José M. Peña and Pilar Herrero

12.1 Introduction 201
12.2 DMGA overview 202
12.3 Horizontal composition 204
12.4 Vertical composition 206
12.5 The need for brokering 208
12.6 Brokering-based data mining grid architecture 209
12.7 Use cases: Apriori, ID3 and J4.8 algorithms 210
   12.7.1 Horizontal composition use case: Apriori 210
   12.7.2 Vertical composition use cases: ID3 and J4.8 213
12.8 Related work 216
12.9 Conclusions 217

13 Grid-based data mining with the Environmental Scenario Search Engine (ESSE) 221
Mikhail Zhizhin, Alexey Poyda, Dmitry Mishin, Dmitry Medvedev, Eric Kihn and Vassily Lyutsarev

13.1 Environmental data source: NCEP/NCAR reanalysis data set 222
13.2 Fuzzy search engine 223
   13.2.1 Operators of fuzzy logic 224
   13.2.2 Fuzzy logic predicates 226
   13.2.3 Fuzzy states in time 227
   13.2.4 Relative importance of parameters 229
   13.2.5 Fuzzy search optimization 229
13.3 Software architecture 231
   13.3.1 Database schema optimization 231
   13.3.2 Data grid layer 233
<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>13.3.3 ESSE data resource</td>
<td>235</td>
</tr>
<tr>
<td>13.3.4 ESSE data processor</td>
<td>235</td>
</tr>
<tr>
<td>13.4 Applications</td>
<td>237</td>
</tr>
<tr>
<td>13.4.1 Global air temperature trends</td>
<td>238</td>
</tr>
<tr>
<td>13.4.2 Statistics of extreme weather events</td>
<td>239</td>
</tr>
<tr>
<td>13.4.3 Atmospheric fronts</td>
<td>239</td>
</tr>
<tr>
<td>13.5 Conclusions</td>
<td>243</td>
</tr>
<tr>
<td>14 Data pre-processing using OGSA-DAI</td>
<td>247</td>
</tr>
<tr>
<td>Martin Swain and Neil P. Chue Hong</td>
<td></td>
</tr>
<tr>
<td>14.1 Introduction</td>
<td>247</td>
</tr>
<tr>
<td>14.2 Data pre-processing for grid-enabled data mining</td>
<td>248</td>
</tr>
<tr>
<td>14.3 Using OGSA-DAI to support data mining applications</td>
<td>248</td>
</tr>
<tr>
<td>14.3.1 OGSA-DAI’s activity framework</td>
<td>249</td>
</tr>
<tr>
<td>14.3.2 OGSA-DAI workflows for data management and pre-processing</td>
<td>253</td>
</tr>
<tr>
<td>14.4 Data pre-processing scenarios in data mining applications</td>
<td>255</td>
</tr>
<tr>
<td>14.4.1 Calculating a data summary</td>
<td>255</td>
</tr>
<tr>
<td>14.4.2 Discovering association rules in protein unfolding simulations</td>
<td>256</td>
</tr>
<tr>
<td>14.4.3 Mining distributed medical databases</td>
<td>257</td>
</tr>
<tr>
<td>14.5 State-of-the-art solutions for grid data management</td>
<td>258</td>
</tr>
<tr>
<td>14.6 Discussion</td>
<td>259</td>
</tr>
<tr>
<td>14.7 Open Issues</td>
<td>259</td>
</tr>
<tr>
<td>14.8 Conclusions</td>
<td>260</td>
</tr>
<tr>
<td>Index</td>
<td>263</td>
</tr>
</tbody>
</table>