Editor Werner Dubitzky

DATA MINING TECHNIQUES in Grid Computing Environments

WILEY-BLACKWELL

Data Mining Techniques in Grid Computing Environments

Editor

Werner Dubitzky

University of Ulster, UK

WILEY-BLACKWELL

A John Wiley & Sons, Ltd., Publication

This edition first published 2008

Wiley-Blackwell is an imprint of John Wiley & Sons, formed by the merger of Wiley's global Scientific, Technical and Medical business with Blackwell Publishing.

Registered office: John Wiley & Sons Ltd, The Atrium, Southern Gate, Chichester, West Sussex, PO19 8SQ, UK

Other Editorial Offices:
9600 Garsington Road, Oxford, OX4 2DQ, UK
111 River Street, Hoboken, NJ 07030-5774, USA

For details of our global editorial offices, for customer services and for information about how to apply for permission to reuse the copyright material in this book please see our website at www.wiley.com/wiley-blackwell

Library of Congress Cataloguing-in-Publication Data

Dubitzky, Werner, 1958-
Data mining techniques in grid computing environments / Werner Dubitzky.
p. cm.
Includes bibliographical references and index.
ISBN 978-0-470-51258-6 (cloth)
1. Data mining. 2. Computational grids (Computer systems) I. Title.
QA76.9.D343D83 2008
004'.36–dc22

2008031720

ISBN: 978 0 470 51258 6

A catalogue record for this book is available from the British Library.

Set in 10/12 pt Times by Thomson Digital, Noida, India
Printed in Singapore by Markono Print Media Pte Ltd

First printing 2008

Preface

Modern organizations across many sectors rely increasingly on computerized information processes and infrastructures. This is particularly true for high-tech and knowledge sectors such as finance, communication, engineering, manufacturing, government, education, medicine, science and technology. As the underlying information systems evolve and become progressively more sophisticated, their users and managers are facing an exponentially growing volume of increasingly complex data, information, and knowledge. Exploring, analyzing and interpreting this information is a challenging task. Besides traditional statistics-based methods, *data mining* is quickly becoming a key technology in addressing the data analysis and interpretation tasks.

Data mining can be viewed as the formulation, analysis, and implementation of an induction process (proceeding from specific data to general patterns) that facilitates the nontrivial extraction of implicit, previously unknown, and potentially useful information from data. Data mining ranges from highly theoretical mathematical work in areas such as statistics, machine learning, knowledge representation and algorithms to systems solutions for problems like fraud detection, modeling of cancer and other complex diseases, network intrusion, information retrieval on the Web and monitoring of grid systems. Data mining techniques are increasingly employed in traditional scientific discovery disciplines, such as biological, medical, biomedical, chemical, physical and social sciences, and a variety of other knowledge industries, such as governments, education, high-tech engineering and process automation. Thus, data mining is playing a highly important role in structuring and shaping future knowledge-based industries and businesses. Effective and efficient management and use of stored data, and in particular the computer-assisted transformation of these data into information and knowledge, is considered a key factor for success.

While the need for sophisticated data mining solutions is growing quickly, it has been realized that conventional data and computer systems and infrastructures are often too limited to meet the requirements of modern data mining applications. Very large data volumes require significant processing power and data throughput. Dedicated and specialized hardware and software are usually tied to particular geographic locations or sites and therefore require the data and data mining tools and programs to be translocated in a flexible, seamless and efficient fashion. Commonly, people and organizations working simultaneously on a large-scale problem tend to reside at geographically dispersed sites, necessitating sophisticated distributed data mining tools and infrastructures. The requirements arising from such large-scale, distributed data mining scenarios are extremely demanding and it is unlikely that a single "killer solution" will emerge that satisfies them all. There is a way forward, though. Two recently emerging computer technologies promise to play a major role in the evolution of future, advanced data mining applications: *grid computing* and *Web services*.

Grid refers to persistent computing environments that enable software applications to integrate processors, storage, networks, instruments, applications and other resources that are managed by diverse organizations in dispersed locations. Web services are broadly regarded as self-contained, self-describing, modular applications that can be published, located, and invoked across the Internet. Recent developments are designed to bring about a convergence of grid and Web services technology (e.g. service-oriented architectures, WSRF). Grid computing and Web services and their future incarnations have a great potential for becoming a fundamental pillar of advanced data mining solutions in science and technology. This volume investigates data mining in the context of grid computing and, to some extent, Web services. In particular, this book presents a detailed account of what motivates the grid-enabling of data mining applications and what is required to develop and deploy such applications. By conveying the experience and lessons learned from the synergy of data mining and grid computing, we believe that similar future efforts could benefit in multiple ways, not least by being able to identify and avoid potential pitfalls and caveats involved in developing and deploying data mining solutions for the grid. We further hope that this volume will foster the understanding and use of grid-enabled data mining technology and that it will help standardization efforts in this field.

The approach taken in this book is conceptual and practical in nature. This means that the presented technologies and methods are described in a largely non-mathematical way, emphasizing data mining tasks, user and system requirements, information processing, IT and system architecture elements. In doing so, we avoid requiring the reader to possess detailed knowledge of advanced data mining theory and mathematics. Importantly, the merits and limitations of the presented technologies and methods are discussed on the basis of real-world case studies.

Our goal in developing this book is to address complex issues arising from grid-enabling data mining applications in different domains, by providing what is simultaneously a *design blueprint*, *user guide*, and *research agenda* for current and future developments in the field.

As *design blueprint*, the book is intended for the practicing professional (analyst, researcher, developer, senior executive) tasked with (a) the analysis and interpretation of large volumes of data requiring the sharing of resources, (b) the grid-enabling of existing data mining applications, and (c) the development and deployment of generic and novel enabling technology in the context of grid computing, Web services and data mining.

As a *user guide*, the book seeks to address the requirements of scientists and researchers to gain a basic understanding of existing concepts, methodologies and systems, combining data mining and modern distributed computing technology. To assist such users, the key concepts and assumptions of the various techniques, their conceptual and computational merits and limitations are explained, and guidelines for choosing the most appropriate technologies are provided.

As a *research agenda*, this volume is intended for students, educators, scientists and research managers seeking to understand the state of the art of data mining in grid computing environments and to identify the areas in which gaps in our knowledge demand further research and development. To this end, our aim is to maintain readability and accessibility throughout the chapters, rather than compiling a mere reference manual. Therefore, considerable effort is made to ensure that the presented material is supplemented by rich literature cross-references to more foundational work and ongoing developments.

Clearly, we cannot expect to do full justice to all three goals in a single book. However, we do believe that this book has the potential to go a long way in fostering the understanding, development and deployment of data mining solutions in grid computing and Web services environments. Thus, we hope this volume will contribute to increased communication and collaboration across various data mining and IT disciplines and will help facilitate a consistent approach to data mining in distributed computing environments in the future.

Acknowledgments

We thank the contributing authors for their contributions and for meeting the stringent deadlines and quality requirements. This work was supported by the European Commission FP6 grants No. 004475 (the DataMiningGrid[1] project), No. 033883 (the QosCosGrid[2] project), and No. 033437 (the Chemomentum[3] project).

Werner Dubitzky

Coleraine and Oberstaufen
March 2008

[1] www.DataMiningGrid.org
[2] www.QosCosGrid.com
[3] http://www.chemomentum.org/

List of Contributors

Andy Brass
University of Manchester
School of Computer Science
Manchester
United Kingdom
andy.brass@manchester.ac.uk

Peter Brezany
Institute of Scientific Computing
University of Vienna
Vienna
Austria
brezany@par.univie.ac.at

Eugenio Cesario
ICAR-CNR
Rende (CS)
Italy
cesario@icar.cnr.it

William K. Cheung
Department of Computer Science
Hong Kong Baptist University
Kowloon Tong
Hong Kong
william@comp.hkbu.edu.hk

Neil P. Chue Hong
EPCC
The University of Edinburgh
Edinburgh
United Kingdom
n.chuehong@omii.ac.uk

Antonio Congiusta
DEIS – University of Calabria and
DIIMA – University of Salerno
Rende (CS) and Salerno
Italy
acongiusta@deis.unical.it

Helen Conover
Information Technology & Systems
Center
The University of Alabama
Huntsville, USA
hconover@itsc.uah.edu

Vasa Curcin
Imperial College London
Department of Computing
London
United Kingdom
vc100@doc.ic.ac.uk

Werner Dubitzky
Biomedical Sciences Research Institute
University of Ulster
Coleraine
United Kingdom
w.dubitzky@ulster.ac.uk

Renato A. Ferreira
Universidade Federal de Minas Gerais
Department of Computer Science
Minas Gerais
Brazil
renato@dcc.ufmg.br

Paul Fisher
University of Manchester
School of Computer Science
Manchester
United Kingdom
pfisher@cs.manchester.ac.uk

Moustafa Ghanem
Imperial College London
Department of Computing
London
United Kingdom
mmg@doc.ic.ac.uk

Carole Goble
University of Manchester
School of Computer Science
Manchester
United Kingdom
carole.goble@manchester.ac.uk

Sara Graves
Information Technology & Systems Center
The University of Alabama
Huntsville, USA
sgraves@itsc.uah.edu

Pierre Gueant
Universidad Politécnica de Madrid
Facultad de Informática
Madrid
Spain
pgueant@fi.upm.es

Dorgival O. Guedes
Universidade Federal de Minas Gerais
Department of Computer Science
Minas Gerais
Brazil
dorgival@dcc.ufmg.br

Yike Guo
Imperial College London
Department of Computing
London
United Kingdom
yg@doc.ic.ac.uk

Pilar Herrero
Universidad Politécnica de Madrid
Facultad de Informática
Madrid
Spain
pherrero@fi.upm.es

Ivan Janciak
Institute of Scientific Computing
University of Vienna
Vienna
Austria
janciak@par.univie.ac.at

Ken Keiser
Information Technology & Systems Center
The University of Alabama
Huntsville, USA
kkeiser@itsc.uah.edu

Eric Kihn
22 National Geophysical Data Center
NOAA
Bolder, CO
USA
eric.a.kihn@noaa.gov

Arie Leizarowitz
Technion -- Israel Institute of Technology
Department of Mathematics
Haifa
Israel
la@techunix.technion.ac.il

Hong Lin
Information Technology & Systems Center
The University of Alabama
Huntsville, USA
alin@itsc.uah.edu

Vassily Lyutsarev
Microsoft Research Cambridge
Microsoft Research Ltd.
Cambridge
United Kingdom
vassilyl@microsoft.com

Manil Maskey
Information Technology & Systems Center
The University of Alabama
Huntsville, USA
mmaskey@itsc.uah.edu

Michael May
Fraunhofer IAIS
Sankt Augustin
Germany
michael.may@iais.fraunhofer.de

Dmitry Medvedev
Geophysical Center RAS
Moscow
Russia
dmedv@wdcb.ru

Wagner Meira Jr.
Universidade Federal de Minas Gerais
Department of Computer Science
Minas Gerais
Brazil
meira@dcc.ufmg.br

Pedro de Miguel
Universidad Politécnica de Madrid
Facultad de Informática
Madrid
Spain
pmiguel@fi.upm.es

Dmitry Mishin
Geophysical Center RAS
Moscow
Russia
dimm@wdcb.ru

Jesús Montes
Universidad Politécnica de Madrid
Facultad de Informática
Madrid
Spain
jmontes@fi.upm.es

Noam Palatin
Technion -- Israel Institute of Technology
Department of Mathematics
Haifa
Israel
noampalatin@gmail.com

José M. Peña
Universidad Politécnica de Madrid
Facultad de Informática
Madrid
Spain
jmpena@fi.upm.es

María S. Pérez
Universidad Politécnica de Madrid
Facultad de Informática
Madrid
Spain
mperez@fi.upm.es

Alexey Poyda
Moscow State University
Moscow
Russia
poyda@wdcb.ru

Rahul Ramachandran
Information Technology & Systems Center
The University of Alabama
Huntsville, USA
rramachandran@itsc.uah.edu

Omer F. Rana
School of Computer Science
Cardiff University
Cardiff
United Kingdom
o.f.rana@cs.cardiff.ac.uk

John Rushing
Information Technology & Systems Center
The University of Alabama
Huntsville, USA
jrushing@itsc.uah.edu

Alberto Sánchez
Universidad Politécnica de Madrid
Facultad de Informática
Madrid
Spain
ascampos@fi.upm.es

Assaf Schuster
Technion -- Israel Institute of Technology
Department of Computer Science
Haifa
Israel
assaf@cs.technion.ac.il

Ali Shaikh Ali
School of Computer Science
Cardiff University
Cardiff
United Kingdom
ali.shaikhali@cs.cardiff.ac.uk

Robert Stevens
University of Manchester
School of Computer Science
Manchester
United Kingdom
robert.stevens@manchester.ac.uk

Martin Swain
Biomedical Sciences Research Institute
University of Ulster
Coleraine
United Kingdom
mt.swain@ulster.ac.uk

Domenico Talia
DEIS – University of Calabria
Rende (CS)
Italy
talia@deis.unical.it

A. Min Tjoa
Institute of Software Technology & Interactive Systems
Vienna University of Technology
Vienna, Austria
tjoa@ifs.tuwien.ac.at

Paolo Trunfio
DEIS – University of Calabria
Rende (CS)
Italy
trunfio@deis.unical.it

Julio J. Valdés
Institute for Information Technology
National Research Council
Ottawa
Canada
Julio.Valdes@nrc-cnrc.gc.ca

Dennis Wegener
Fraunhofer IAIS
Sankt Augustin
Germany
dennis.wegener@iais.fraunhofer.de

Patrick Wendel
InforSense Ltd.
London
United Kingdom
patrick@inforsense.com

Ran Wolff
Technion -– Israel Institute of Technology
Department of Computer Science
Haifa
Israel
rwolff@mis.haifa.ac.il

Jun Zhao
University of Manchester
School of Computer Science
Manchester
United Kingdom
jun.zhao@zoo.ox.ac.uk

Mikhail Zhizhin
Geophysical Center RAS
Moscow
Russia
jjn@wdcb.ru

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 **37**
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 **57**
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 **71**
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

5.4 Data analysis 77
5.4.1 General approach 77
5.4.2 Notation 78
5.4.3 Algorithm 78
5.4.4 Correctness and termination 80
5.5 The GMS 80
5.6 Evaluation 82
5.6.1 Qualitative results 82
5.6.2 Quantitative results 83
5.6.3 Interoperability 85
5.7 Conclusions and future work 88

6 FAEHIM: Federated Analysis Environment for Heterogeneous Intelligent Mining 91
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 105
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

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** **141**
Robert Stevens, Paul Fisher, Jun Zhao, Carole Goble and Andy Brass

9.1 Introduction 141
9.2 The bioinformatics data landscape 143
9.3 The bioinformatics experiment landscape 143
9.4 Taverna for bioinformatics experiments 145
9.4.1 Three-tiered enactment in Taverna 146
9.4.2 The open-typing data models 147
9.5 Building workflows in Taverna 148
9.5.1 Designing a SCUFL workflow 149
9.6 Workflow case study 150
9.6.1 The bioinformatics task 152
9.6.2 Current approaches and issues 153
9.6.3 Constructing workflows 154
9.6.4 Candidate genes involved in trypanosomiasis resistance 156
9.6.5 Workflows and the systematic approach 157
9.7 Discussion 159

10 **Specification of distributed data mining workflows with DataMiningGrid** **165**
Dennis Wegener and Michael May

10.1 Introduction 165
10.2 DataMiningGrid environment 167
10.2.1 General architecture 167
10.2.2 Grid environment 167
10.2.3 Scalability 167
10.2.4 Workflow environment 168
10.3 Operations for workflow construction 169
10.3.1 Chaining 169
10.3.2 Looping 169
10.3.3 Branching 170
10.3.4 Shipping algorithms 170
10.3.5 Shipping data 170
10.3.6 Parameter variation 171
10.3.7 Parallelization 171
10.4 Extensibility 171
10.5 Case studies 173
10.5.1 Evaluation criteria and experimental methodology 173
10.5.2 Partitioning data 173
10.5.3 Classifier comparison scenario 175
10.5.4 Parameter optimization 175
10.6 Discussion and related work 175
10.7 Open issues 176
10.8 Conclusions 176

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