

## Contents

|                                                                                                         |             |
|---------------------------------------------------------------------------------------------------------|-------------|
| <b>PREFACE</b>                                                                                          | <b>xi</b>   |
| <b>FOREWORD</b>                                                                                         | <b>xiii</b> |
| <b>LIST OF CONTRIBUTORS</b>                                                                             | <b>xv</b>   |
| <b>Chapter 1    INFORMATION FILTERS SUPPLYING DATA<br/>WAREHOUSES WITH BENCHMARKING<br/>INFORMATION</b> | <b>1</b>    |
| <i>Witold Abramowicz, Paweł Jan Kalczyński, Krzysztof Węcel</i>                                         |             |
| 1. Introduction                                                                                         | 1           |
| 2. Data Warehouses                                                                                      | 2           |
| 3. The HyperSDI System                                                                                  | 4           |
| 4. User Profiles in the HyperSDI System                                                                 | 11          |
| 5. Building Data Warehouse Profiles                                                                     | 11          |
| 6. Techniques for Improving Profiles                                                                    | 18          |
| 7. Implementation Notes                                                                                 | 22          |
| 8. Conclusions                                                                                          | 25          |
| References                                                                                              |             |
| <b>Chapter 2    PARALLEL MINING OF ASSOCIATION RULES</b>                                                | <b>29</b>   |
| <i>David Cheung, Sau Dan Lee</i>                                                                        |             |
| 1. Introduction                                                                                         | 29          |
| 2. Parallel Mining of Association Rules                                                                 | 32          |
| 3. Pruning Techniques and The FPM Algorithm                                                             | 33          |
| 4. Metrics for Data Skewness and Workload Balance                                                       | 39          |
| 5. Partitioning of the Database                                                                         | 48          |
| 6. Experimental Evaluation of the Partitioning Algorithms                                               | 56          |
| 7. Discussions                                                                                          | 62          |
| 8. Conclusions                                                                                          | 64          |
| References                                                                                              | 65          |

|                      |                                                                                                                         |                |
|----------------------|-------------------------------------------------------------------------------------------------------------------------|----------------|
| <b>Chapter 3</b>     | <b>UNSUPERVISED FEATURE RANKING<br/>AND SELECTION</b>                                                                   | <b>67</b>      |
|                      | <i>Manoranjan Dash, Huan Liu, Jun Yao</i>                                                                               |                |
|                      | 1. Introduction                                                                                                         | 67             |
|                      | 2. Basic Concepts and Possible Approaches                                                                               | 69             |
|                      | 3. An Entropy Measure for Continuous and Nominal Data Types                                                             | 72             |
|                      | 4. Algorithm to Find Important Variables                                                                                | 75             |
|                      | 5. Experimental Studies                                                                                                 | 76             |
|                      | 6. Clustering Using SUD                                                                                                 | 80             |
|                      | 7. Discussion and Conclusion                                                                                            | 82             |
|                      | References                                                                                                              | 84             |
| <br><b>Chapter 4</b> | <br><b>APPROACHES TO CONCEPT BASED EXPLORATION<br/>OF INFORMATION RESOURCES</b>                                         | <br><b>89</b>  |
|                      | <i>Hele-Mai Haav, Jørgen Fischer Nilsson</i>                                                                            |                |
|                      | 1. Introduction                                                                                                         | 89             |
|                      | 2. Conceptual Taxonomies                                                                                                | 91             |
|                      | 3. Ontology Driven Concept Retrieval                                                                                    | 99             |
|                      | 4. Search based on formal concept analysis                                                                              | 104            |
|                      | 5. Conclusion                                                                                                           | 109            |
|                      | Acknowledgements                                                                                                        | 109            |
|                      | References                                                                                                              | 109            |
| <br><b>Chapter 5</b> | <br><b>HYBRID METHODOLOGY OF KNOWLEDGE<br/>DISCOVERY FOR BUSINESS INFORMATION</b>                                       | <br><b>111</b> |
|                      | <i>Zdzisław S. Hippe</i>                                                                                                |                |
|                      | 1. Introduction                                                                                                         | 111            |
|                      | 2. Present Status of Data Mining                                                                                        | 113            |
|                      | 3. Experiments with Mining Regularities from Data                                                                       | 118            |
|                      | 4. Discussion                                                                                                           | 125            |
|                      | Acknowledgements                                                                                                        | 126            |
|                      | References                                                                                                              | 126            |
| <br><b>Chapter 6</b> | <br><b>FUZZY LINGUISTIC SUMMARIES OF DATABASES<br/>FOR AN EFFICIENT BUSINESS DATA ANALYSIS<br/>AND DECISION SUPPORT</b> | <br><b>129</b> |
|                      | <i>Janusz Kacprzyk, Ronald R. Yager and Sławomir Zadrozny</i>                                                           |                |
|                      | 1. Introduction                                                                                                         | 129            |
|                      | 2. Idea of Linguistic Summaries Using Fuzzy Logic with<br>Linguistic Quantifiers                                        | 131            |
|                      | 3. On Other Validity Criteria                                                                                           | 134            |
|                      | 4. Derivation of Linguistic Summaries via a Fuzzy Logic<br>Based Database Querying Interface                            | 140            |
|                      | 5. Implementation for a Sales Database at a Computer Retailer                                                           | 147            |

|                                                                                              |            |
|----------------------------------------------------------------------------------------------|------------|
| 6. Concluding Remarks                                                                        | 150        |
| References                                                                                   | 150        |
| <b>Chapter 7 INTEGRATING DATA SOURCES USING<br/>A STANDARDIZED GLOBAL DICTIONARY</b>         | <b>153</b> |
| <i>Ramon Lawrence and Ken Barker</i>                                                         |            |
| 1. Introduction                                                                              | 154        |
| 2. Data Semantics and the Integration Problem                                                | 154        |
| 3. Previous work                                                                             | 156        |
| 4. The Integration Architecture                                                              | 157        |
| 5. The Global Dictionary                                                                     | 160        |
| 6. The Relational Integration Model                                                          | 164        |
| 7. Special Cases of Integration                                                              | 169        |
| 8. Applications to the WWW                                                                   | 171        |
| 9. Future Work and Conclusions                                                               | 171        |
| References                                                                                   | 172        |
| <b>Chapter 8 MAINTENANCE OF DISCOVERED ASSOCIATION<br/>RULES</b>                             | <b>173</b> |
| <i>Sau Dan Lee, David Cheung</i>                                                             |            |
| 1. Introduction                                                                              | 173        |
| 2. Problem Description                                                                       | 176        |
| 3. The FUP Algorithm for the Insertion Only Case                                             | 179        |
| 4. The FUP Algorithm for the Deletions Only Case                                             | 183        |
| 5. The FUP <sub>2</sub> Algorithm for the General Case                                       | 189        |
| 6. Performance Studies                                                                       | 194        |
| 7. Discussions                                                                               | 204        |
| 8. Conclusions                                                                               | 208        |
| Notes                                                                                        | 209        |
| References                                                                                   | 209        |
| <b>Chapter 9 MULTIDIMENSIONAL BUSINESS PROCESS<br/>ANALYSIS WITH THE PROCESS WAREHOUSE</b>   | <b>211</b> |
| <i>Beate List, Josef Schiefer, A Min Tjoa, Gerald Quirchmayr</i>                             |            |
| 1. Introduction                                                                              | 211        |
| 2. Related Work                                                                              | 213        |
| 3. Goals of the Data Warehouse Approach                                                      | 215        |
| 4. Data Source                                                                               | 216        |
| 5. Basic Process Warehouse Components Representing<br>Business Process Analysis Requirements | 216        |
| 6. Data Model and Analysis Capabilities                                                      | 219        |
| 7. Conclusion and Further Research                                                           | 225        |
| References                                                                                   | 225        |

|                   |                                                                                                                |            |
|-------------------|----------------------------------------------------------------------------------------------------------------|------------|
| <b>Chapter 10</b> | <b>AMALGAMATION OF STATISTICS AND DATA MINING TECHNIQUES: EXPLORATIONS IN CUSTOMER LIFETIME VALUE MODELING</b> | <b>229</b> |
|                   | <i>D. R. Mani, James Drew, Andrew Betz and Piew Datta</i>                                                      |            |
|                   | 1. Introduction                                                                                                | 229        |
|                   | 2. Statistics and Data Mining Techniques: A Characterization                                                   | 231        |
|                   | 3. Lifetime Value (LTV) Modeling                                                                               | 232        |
|                   | 4. Customer Data for LTV Tenure Prediction                                                                     | 234        |
|                   | 5. Classical Statistical Approaches to Survival Analysis                                                       | 235        |
|                   | 6. Neural Networks for Survival Analysis                                                                       | 239        |
|                   | 7. From Data Models to Business Insight                                                                        | 244        |
|                   | 8. Conclusion: The Amalgamation of Statistical and Data Mining Techniques                                      | 247        |
|                   | References                                                                                                     | 249        |
| <b>Chapter 11</b> | <b>ROBUST BUSINESS INTELLIGENCE SOLUTIONS</b>                                                                  | <b>251</b> |
|                   | <i>Jan Mrazek</i>                                                                                              |            |
|                   | 1. Introduction                                                                                                | 251        |
|                   | 2. Business Intelligence Architecture                                                                          | 252        |
|                   | 3. Data Transformation                                                                                         | 258        |
|                   | 4. Data Modelling                                                                                              | 260        |
|                   | 5. Integration Of Data Mining                                                                                  | 268        |
|                   | 6. Conclusion                                                                                                  | 272        |
|                   | References                                                                                                     | 273        |
| <b>Chapter 12</b> | <b>THE ROLE OF GRANULAR INFORMATION IN KNOWLEDGE DISCOVERY IN DATABASES</b>                                    | <b>275</b> |
|                   | <i>Witold Pedrycz</i>                                                                                          |            |
|                   | 1. Introduction                                                                                                | 276        |
|                   | 2. Granulation of information                                                                                  | 277        |
|                   | 3. The development of data-justifiable information granules                                                    | 284        |
|                   | 4. Building associations in databases                                                                          | 287        |
|                   | 5. From associations to rules in databases                                                                     | 292        |
|                   | 6. The construction of rules in data mining                                                                    | 293        |
|                   | 7. Properties of rules induced by associations                                                                 | 298        |
|                   | 8. Detailed computations of the consistency of rules and its analysis                                          | 300        |
|                   | 9. Conclusions                                                                                                 | 303        |
|                   | Acknowledgment                                                                                                 | 304        |
|                   | References                                                                                                     | 304        |

|                       |                                                                                                                                              |                |
|-----------------------|----------------------------------------------------------------------------------------------------------------------------------------------|----------------|
| <b>Chapter 13</b>     | <b>DEALING WITH DIMENSIONS IN DATA<br/>WAREHOUSING</b>                                                                                       | <b>307</b>     |
|                       | <i>Jaroslav Pokorny</i>                                                                                                                      |                |
|                       | 1. Introduction                                                                                                                              | 308            |
|                       | 2. DW Modelling with Tables                                                                                                                  | 310            |
|                       | 3. Dimensions                                                                                                                                | 311            |
|                       | 4. Constellations                                                                                                                            | 314            |
|                       | 5. Dimension Hierarchies with ISA-hierarchies                                                                                                | 315            |
|                       | 6. Conclusions                                                                                                                               | 323            |
|                       | References                                                                                                                                   | 324            |
| <br><b>Chapter 14</b> | <br><b>ENHANCING THE KDD PROCESS IN THE<br/>RELATIONAL DATABASE MINING FRAMEWORK BY<br/>QUANTITATIVE EVALUATION OF ASSOCIATION<br/>RULES</b> | <br><b>325</b> |
|                       | <i>Giuseppe Psaila</i>                                                                                                                       |                |
|                       | 1. Introduction                                                                                                                              | 325            |
|                       | 2. The Relational Database Mining Framework                                                                                                  | 327            |
|                       | 3. The Evaluate Rule Operator                                                                                                                | 331            |
|                       | 4. Enhancing the Knowledge Discovery Process                                                                                                 | 344            |
|                       | 5. Conclusions and Future Work                                                                                                               | 348            |
|                       | Notes                                                                                                                                        | 349            |
|                       | References                                                                                                                                   | 349            |
| <br><b>Chapter 15</b> | <br><b>SPEEDING UP HYPOTHESIS DEVELOPMENT</b>                                                                                                | <br><b>351</b> |
|                       | <i>Jörg A. Schlösser, Peter C. Lockemann, Matthias Gimbel</i>                                                                                |                |
|                       | 1. Introduction                                                                                                                              | 352            |
|                       | 2. Information Model                                                                                                                         | 354            |
|                       | 3. The Execution Architecture of CITRUS                                                                                                      | 356            |
|                       | 4. Searching the Information Directory                                                                                                       | 359            |
|                       | 5. Documentation of the Process History                                                                                                      | 361            |
|                       | 6. Linking the Information Model with the Relational Model                                                                                   | 362            |
|                       | 7. Generation of SQL Queries                                                                                                                 | 364            |
|                       | 8. Automatic Materialization of Intermediate Results                                                                                         | 368            |
|                       | 9. Experimental Results                                                                                                                      | 369            |
|                       | 10. Utilizing Past Experience                                                                                                                | 371            |
|                       | 11. Related Work                                                                                                                             | 372            |
|                       | 12. Concluding Remarks                                                                                                                       | 374            |
|                       | References                                                                                                                                   | 374            |

|                       |                                                                                                                                                                     |                                        |
|-----------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------|----------------------------------------|
| <b>Chapter 16</b>     | <b>SEQUENCE MINING IN DYNAMIC<br/>AND INTERACTIVE ENVIRONMENTS</b>                                                                                                  | <b>377</b>                             |
|                       | <i>Srinivasan Parthasarathy, Mohammed J. Zaki, Mitsunori Ogihara,<br/>Sandhya Dwarkadas</i>                                                                         |                                        |
| 1.                    | Introduction                                                                                                                                                        | 378                                    |
| 2.                    | Problem Formulation                                                                                                                                                 | 379                                    |
| 3.                    | The SPADE Algorithm                                                                                                                                                 | 382                                    |
| 4.                    | Incremental Mining Algorithm                                                                                                                                        | 384                                    |
| 5.                    | Interactive Sequence Mining                                                                                                                                         | 388                                    |
| 6.                    | Experimental Evaluation                                                                                                                                             | 390                                    |
| 7.                    | Related Work                                                                                                                                                        | 394                                    |
| 8.                    | Conclusions                                                                                                                                                         | 395                                    |
|                       | Acknowledgements                                                                                                                                                    | 395                                    |
|                       | References                                                                                                                                                          | 395                                    |
| <br><b>Chapter 17</b> | <br><b>INVESTIGATION OF ARTIFICIAL NEURAL<br/>NETWORKS FOR CLASSIFYING LEVELS<br/>OF FINANCIAL DISTRESS OF FIRMS: THE CASE<br/>OF AN UNBALANCED TRAINING SAMPLE</b> | <br><br><br><br><br><br><br><b>397</b> |
|                       | <i>Jozef Zurada, Benjamin P. Foster, Terry J. Ward</i>                                                                                                              |                                        |
| 1.                    | Introduction                                                                                                                                                        | 398                                    |
| 2.                    | Motivation and Literature Review                                                                                                                                    | 399                                    |
| 3.                    | Logit Regression, Neural Network, and Principal<br>Component Analysis Fundamentals                                                                                  | 402                                    |
| 4.                    | Research Methodology                                                                                                                                                | 409                                    |
| 5.                    | Discussion of the Results                                                                                                                                           | 415                                    |
| 6.                    | Conclusions and Future Research Directions                                                                                                                          | 420                                    |
|                       | Appendix - Neural Network Toolbox                                                                                                                                   | 421                                    |
|                       | References                                                                                                                                                          | 423                                    |
| <br><b>INDEX</b>      |                                                                                                                                                                     | <br><b>425</b>                         |