Preface

Nowadays, the data management community acknowledges the fact that data are not only numerical or symbolic, but that they may be:

- represented in various formats (databases, texts, images, sounds, videos, etc.);
- diversely structured (relational databases, XML document repositories, etc.);
- originating from several different sources (distributed databases, the Web, etc.);
- described through several channels or points of view (radiographies and audio diagnosis of a physician, data expressed in different scales or languages, etc.); and
- changing in terms of definition or value (temporal databases, periodical surveys, etc.).

Data that fall in several of the above categories may be termed as complex data (Darmont et al., 2005). Managing such data involves a lot of different issues regarding their structure, storage and processing. However, in many decision-support fields (customer relationship management, marketing, competition monitoring, medicine, etc.), they are the real data that need to be exploited.

The advent of complex data indeed imposes another vision of decision-support processes such as data warehousing and data mining. The classic architectures of data warehouses (Inmon, 2002; Kimball & Ross, 2002) have shown their efficiency when data are "simple" (i.e., numerical or symbolic). However, these architectures must be completely reconsidered when dealing with complex data.

For instance, the concept of a centralized warehouse might not be pertinent in all cases. Indeed, the specificity of complex data and their physical location rather impose new solutions based on virtual warehousing or architecture-oriented approaches. Data integration through mediation systems may also be considered as a new approach for the ETL (extract, transform, load) process (Kimball & Caserta, 2004). Furthermore, online analytical processing, better known as OLAP (Thomsen, 2002), must surpass its initial vocation to allow more effective analyses. The combination of OLAP and data mining techniques is a new challenge imposed by complex data.

As for data mining techniques, they generally cannot apply directly onto complex data. Usual data representation spaces for classical data mining algorithms (Hand, Mannila, & Smyth, 2001; Witten & Frank, 2005) are not adapted. Either it is necessary to perform an important, often intricate, preprocessing work to map complex data into these representation spaces without losing information (Pyle, 1999), or it is necessary to substantially modify data-mining algorithms to take into account the specificity of complex data.

The complex data research topic is currently just starting to emerge. Though many people actually work on subsets of complex data, such as multimedia data, the idea of a broader field is just starting to spread. This book is designed to provide readers with an overall view of this emerging field of complex data processing by bringing together various research studies and surveys in different subfields, and by highlighting the similarities between the different data, issues and approaches. It is expected that researchers in universities and research institutions will find such discussions particularly insightful and helpful to their current and future research. In addition, this book is also designed to serve technical professionals, since many existing applications could benefit from the exploitation of other types of data than the ones they usually draw on.

This book is organized into two major sections dealing respectively with complex data warehousing (including spatial, XML and text warehousing) and complex data mining (including distance metrics and similarity measures, pattern management, multimedia and gene sequence mining).

Section I: Complex Data Warehousing

Chapter I: Spatial Data Warehouse Modelling, by Damiani and Spaccapietra, is concerned with multidimensional data models for spatial data warehouses. It first draws a picture of the research area, and then introduces a novel spatial multidimensional data model for spatial objects with geometry: the Multigranular Spatial Data warehouse (MuSD). The main novelty of the model is the representation of spatial measures at multiple levels of geometric granularity.

Chapter II: Goal-Oriented Requirement Engineering for XML Document Warehouses, by Nassis et al. discusses the need of capturing data warehouse requirements early in the design stage, and explores a requirement engineering approach, namely the goal-oriented approach. This approach is then extended to introduce the XML document warehouse (XDW) requirement model.

Chapter III: Building an Active Content Warehouse, by Abiteboul, Nguyen and Ruberg, introduces the concept of content warehousing: the management of loosely structured data. The construction and maintenance of a content warehouse is an intricate task, so the authors propose the Acware (active content warehouse) specification language to help all sorts of users to organize content in a simple manner. This approach is based on XML and Web Services.

Chapter IV: Text Warehousing: Present and Future, by Badia, is part overview of document warehouse and information retrieval techniques, part position paper. The author introduces a new paradigm, based in information extraction, for true integration, and analyzes the challenges that stand in the way of this technology being widely used. He also discusses some considerations on future developments in the general area of documents in databases.

Chapter V: Morphology, Processing, and Integrating of Information from Large Source Code Warehouses for Decision Support, by Rech, describes the morphology of object-oriented source code and how it is processed, integrated and used for knowledge discovery in software engineering in order to support decision-making regarding the refactoring, reengineering and reuse of software systems.

Chapter VI: Managing Metadata in Decision Environments, by Shankaranarayanan and Even, describes the implications for managing metadata, which is a key factor for the successful implementation of complex decision environments. Crucial gaps for integrating metadata are identified by comparing the requirements for managing metadata with the capabilities offered by commercial software products. The authors then propose a conceptual architecture for the design of an integrated metadata repository that attempts to redress these gaps.

Chapter VII: DWFIST: The Data Warehouse of Frequent Itemsets Tactics Approach, by Monteiro et al. presents the core of the DWFIST approach, which is concerned with supporting the analysis and exploration of frequent itemsets and derived patterns such as association rules in transactional datasets. The goal of this new approach is to provide flexible pattern-retrieval capabili-

ties without requiring the original data during the analysis phase, and a standard modeling for data warehouses of frequent itemsets allowing an easier development and reuse of tools for analysis and exploration of itemset-based patterns.

Section II: Complex Data Mining

Chapter VIII: On the Usage of Structural Distance Metrics for Mining Hierarchical Structures, by Dalamagas, Cheng and Sellis, studies distance metrics that capture the structural similarity between hierarchical structures and approaches that exploit structural distance metrics to perform mining tasks on hierarchical structures, especially XML documents.

Chapter IX: Structural Similarity Measures in Sources of XML Documents, by Guerrini, Mesiti and Bertino, discusses existing approaches to evaluate and measure structural similarity in sources of XML documents. The most relevant applications of such measures, discussed throughout the chapter, are for document classification, schema extraction, and for document and schema structural clustering.

Chapter X: Pattern Management: Practice and Challenges, by Catania and Maddalena, provides a critical comparison of the existing approaches for pattern management. In particular, specific issues concerning pattern management systems, pattern models and pattern languages are discussed. Several parameters are also identified and used in evaluating the effectiveness of theoretical and industrial proposals.

Chapter XI: VRMiner: A Tool for Multimedia Database Mining with Virtual Reality, by Azzag et al. presents a new 3-D interactive method for visualizing multimedia data with a virtual reality named VRMiner. Navigating through the data is done in a very intuitive and precise way with a 3-D sensor that simulates a virtual camera. Interactive requests can be formulated by the expert with a data glove that recognizes hand gestures. The authors illustrate how this tool has been successfully applied to several real world applications.

Chapter XII: Mining in Music Databases, by Karydis, Nanopoulos and Manolopoulos provides a broad survey of music data mining, including clustering, classification and pattern discovery in music. Throughout the chapter, practical applications of music data mining are presented. This chapter encapsulates the theory and methods required in order to discover knowledge in the form of patterns for music analysis and retrieval, or statistical models for music classification and generation.

Finally, **Chapter XIII: Data Mining in Gene Expression Data Analysis: A Survey**, by Han, Gruenwald and Conway, surveys data mining techniques that have been used for clustering, classification and association rules for gene expression data analysis. In addition, the authors provide a comprehensive list of currently available commercial and academic data mining software together with their features, and finally suggest future research directions.

By gathering this collection of papers of high scientific quality, our main objective is to contribute to the emergence of already existing research in the complex data field. One ambition of this book is to become one of the first foundation references in this new and ambitious research field. We hope it will succeed.

Jérôme Darmont and Omar Boussaïd

Lyon, France

January 2006

References

Darmont, J., Boussaïd, O., Ralaivao, J. C., & Aouiche, K. (2005). An architecture framework for complex data warehouses. *Seventh International Conference on Enterprise Information Systems ICEIS '05*, Miami, FL (pp. 370-373).

Hand, D. J., Mannila, H., & Smyth, P. (2001). *Principles of data mining*. Boston: MIT Press.

Inmon, W. H. (2002). *Building the data warehouse* (2nd ed.). New York: John Wiley & Sons.

Kimball, R., & Ross, M. (2002). *The data warehouse toolkit: The complete guide to dimensional modeling* (2nd ed.). New York: John Wiley & Sons.

Kimball, R., & Caserta, J. (2004). *The data warehouse ETL toolkit*. New York: John Wiley & Sons.

Pyle, D. (1999). *Data preparation for data mining*. San Francisco: Morgan Kaufmann.

Thomsen, E. (2002). *OLAP solutions: Building multidimensional information systems* (2nd ed.). New York: John Wiley & Sons.

Witten, I. H., & Frank, E. (2005). *Data mining: Practical machine learning tools and techniques* (2nd edition). San Francisco: Morgan Kaufmann.