Preface

Data warehousing and data mining are related technologies which have seen a significant boost in the last decades, in a way that many of their concepts and techniques have reached a significant level of maturity. They are applied today in most fields of human activity, from commercial to scientific or industrial areas. Today, decision support, data mining, trend analysis and pattern discovery have a large impact on businesses and science alike. This has led to the development of new solutions and approaches, some of them being incorporated into commercial tools and systems, others producing new advancement opportunities in many fields of human knowledge.

Given this evolution, it is important to understand advances that happened in those technologies concerning solutions and applications, how data warehousing and data mining technologies operate and their positive effects on many areas of human activity and knowledge. Moreover, it is very interesting to look at current developments in the underlying technologies and what research opportunities lay ahead.

The two concepts of data warehousing and data mining are in fact very much related to each other, and both research and commercial application areas need to deal with both. Warehousing refers to a multidimensional data organization, its loading, storage and analysis using typical operations that are frequently denoted as "Online Analytical Processing (OLAP)". Data mining, on the other hand, applies certain classes of algorithms to search and discover new knowledge automatically from multidimensional data sets. This means that not only data mining and data warehousing assume a similar base model, as they are related to each other in another way: data warehousing approaches are useful for basic data organization and analysis, while data mining approaches extend this to include further analysis capabilities, by applying certain objective-directed algorithms for finding new knowledge from the data sets.

This book brings together a set of papers discussing current issues in evolving application domains of data warehousing and mining, showing the trends and solutions that are currently being researched and applied concerning foundations and applications of those technologies. One important objective is to look at research results concerning actual application of the technologies, besides their foundations.

The subjects discussed along the book are relevant for both practitioners and scholars. On the practical application side, the reader will find answers to practical issues regarding how these technologies work and are to be applied, as well as cases of applications of the technologies in different fields of knowledge. The researcher and scholar, on the other hand, will have the opportunity to understand the current state-of-the-art and research-related developments and hot topics. The book also serves as a reference for advanced courses on data warehousing and mining, since it discusses state-of-the-art and advances in various relevant issues of the technologies and of their application to real-world problems.

We have been careful in the choice of chapters, by taking into consideration not only the precious input received by the team of reviewers and the quality of the chapters, but also by evaluating the importance of the subject and by structuring a book that reviews and provides new insight into some of the most interesting aspects of these technologies. One particularly important aspect of the process was to provide adequate feedback and help to the authors, and to eliminate many chapters that were not sound enough in some respect.

In the rest of the Preface we introduce the structure and contents of the book, in order to give the reader a roadmap into what is the content of sections and chapters in the book. This book is structured in such a way that readers with different objectives can all find their way directly to the parts that are most interesting to them. For instance, while a student may wish to read the book sequentially from the start to the end of it, readers specially interested in data mining may jump straight ahead into the section on data mining, and then to the section on applications of data warehousing, since those applications also include interesting parts on mining from data warehousing contexts.

STRUCTURE OF THE BOOK

The book addresses both foundational and application issues in data warehousing and mining. It presents both current state-of-the-art and research in infrastructural aspects and application areas. Trends and solutions in different domains are identified and discussed in the chapters. In the following we provide first a list of the sections and an overview of the contents of those sections. Then we provide a summary of each chapter within each section. This section is therefore a precious roadmap into the contents of the book.

The book is divided into three sections, addressing the following concerns:

- Foundation Issues in Data Warehousing and OLAP
- Application Issues and Trends in Data Warehousing and OLAP
- Foundations and Applications of Data Mining and Data Analysis

Section 1 addresses foundations of data warehouses and OLAP. This section features four chapters covering current topics that include data warehouse architectures, modeling of warehouse and OLAP applications and, from a data warehouse organization perspective, management of multiple versions and compression schemes.

After addressing an important set of foundational issues in data warehousing and OLAP in the first section, section 2 proceeds with both application issues and current research trends. It covers very different application domains, including data streams, sensor data, genomics and geographical data warehouses. These are useful both for they insight into state-of-the-art in their respective domains and as research trends. Reflecting the complementarities between warehousing and mining, data mining is also present in most of the papers from this section, although in the context of warehousing and mining.

Section 3 discusses both foundations and applications of data mining technologies. The section starts with an excellent survey on data clustering, and also includes state-of-the-art information on exception mining, social network mining and risk assessment, a data analysis task. In what concerns applications, the section discusses and presents results in areas such as current applications of clustering, stock market surveillance, protein folding, social networks and risk assessment in geostatistics.

Section 1. Foundation Issues in Data Warehousing and OLAP

Chapter 1, *Data Warehouse Architectures: Practices and Trends*, by Xuegang Huang, is a high-level introductory chapter, discussing the concepts behind data warehouse architectures and past and present trends in data warehouse architectures. It considers real-world requirements for data warehouse solutions, and discusses which architectural patterns should be used to solve those requirements. It further describes how the concept of service-orientation may influence future data warehouse architectures and solutions as well.

Chapter 2, *Improving Expressive Power in Modeling Data Warehouse and OLAP Applications*, by Elzbieta Malinowski, discusses how the conceptual multidimensional model can be used to facilitate the representation of complex hierarchies and different kinds of dimensions in comparison to their representation in a relational model and commercial OLAP tools. This chapter is in itself an excellent reference on multidimensional modeling, representation and implementation issues.

Chapter 3, *From Conventional to Multiversion Data Warehouse: Practical Issues*, by Khurram Shahzad, concerns data warehouse versioning. Versioning is quite important in real-world projects, since operational sources or the data warehouse structure itself may evolve. Conventional data warehouses are not prepared to handle these modifications. The chapter, while not a comprehensive survey on the subject, takes a very practical perspective, collecting and integrating concepts, issues and solutions of multiversion data warehouses in a tutorial-like approach, to provide a unified source for users that need to understand version functionality and mechanisms.

Chapter 4, *Compression Schemes of High Dimensional Data for MOLAP*, by K. M. Azharul Hasan, surveys data compression techniques relevant to multidimensional OLAP and discusses important quality issues of MOLAP compression and of existing techniques. Compression is indeed an important issue faced in implementations of data warehouses and in particular for multidimensional OLAP, due to possibly huge size and sparsity of MOLAP representations.

Section 2. Application Issues and Trends in Data Warehousing and OLAP

Chapter 5, *View Management Techniques and Their Application to Data Stream Management,* by Christoph Quix et al., is a very interesting and insightful chapter on the subjects of view management and data stream management, starting with a suggestion that data stream processing shares many similarities with view management in data warehousing. The chapter provides an overview of view maintenance and view selection methods, explains the fundamental issues of data stream management, and discusses how view management techniques from data warehousing are related to data stream management. Finally, it provides directions for future research in view management, data streams, and data warehousing.

Chapter 6, *A Framework for Data Warehousing and Mining in Sensor Stream Application Domains,* by Nan Jiang provides insight into how data collected from sensor devices can be fed into data warehouses and mined. This is a relevant subject, since sensors are increasingly used in many different applications, from weather and environmental monitoring to hospital and factory operation sites, traffic monitoring and so on. The chapter presents a general framework for domain-driven mining of sensor stream applications, and evaluates the proposed framework with experiments on traffic management and environmental monitoring.

Chapter 7, *A Data Warehousing Approach for Genomics Data Meta-Analysis*, by Martine Collard et al., takes a very different application domain, genomics, and shows how data warehousing and mining are relevant in that context. Since experimental micro-array data and sources of biological knowledge are now available on public repositories, comparative analyses involving several experiments become conceivable and hold potentially relevant knowledge. However, manually navigating and searching for similar tendencies in such huge spaces is impracticable. In this context, the authors propose a semantic data warehousing solution based on semantic web technologies that allows to monitoring both the diversity and the volume of all related data.

Chapter 8, *A Multidimensional Model for Correct Aggregation of Geographic Measures*, by Sandro Bimonte et al., discusses aggregation issues in multidimensional, geostatistic, GIS and Spatial OLAP models. The chapter provides a good review of those models and discusses why the multidimensional aggregation of geographic objects (geographic measures) exhibits theoretical and implementation problems. The authors then proceed to propose a solution to that problem within a GeoCube multidimensional model.

Section 3. Foundations and Applications of Data Mining and Data Analysis

Chapter 9, *Novel Trends in Clustering*, by Claudia Plant and Christian Böhm, is a very interesting work featuring state-of-the-art and current trends analysis on data clustering. It is useful for both researchers and practitioners, providing an overview on emerging trends in clustering, including subspace and projected clustering, correlation clustering, semi-supervised clustering, spectral clustering and parameter-free clustering. Requirements from concrete example applications in life sciences and the web provide motivation for the discussion of novel approaches to clustering in this chapter.

Chapter 10, *Recent Advances of Exception Mining in Stock Market*, by Chao Luo et al., offers a survey of current advanced technologies for exception mining in stock markets. Additionally, it proposes and analyses improved approaches for exception mining and discusses future research directions and related issues.

Chapter 11, *Analysis of Content Popularity in Social Bookmarking Systems,* by Symeon Papadopoulos et al., embraces a very current issue of social networks and social bookmarking systems. Modern SBS-based applications permit their users to submit their preferred content, comment on and rate the content of other users and establish social relations with each other. The chapter provides a unified treatment of the phenomenon by studying four aspects of popularity of socially bookmarked content: (a) the distributional properties of content consumption, (b) its evolution in time, (c) the correlation between the semantics of online content and its popularity, and (d) the impact of online social networks on the content consumption behavior of individuals.

Chapter 12, *Using Data Mining Techniques to Probe the Role of Hydrophobic Residues in Protein Folding and Unfolding Simulations,* by Catarina Silva et. al. shows how data mining approaches, in this case hierarchical clustering and association rules mining, is useful in molecular dynamics simulation experiments related to the study of protein folding problem. The protein folding problem is the identification of rules that determine the acquisition of the native, functional, three-dimensional structure of a protein from its linear sequence of amino-acids. Its importance stems from the fact that functional properties of proteins can frequently be related to protein conformation issues, and data mining methods – hierarchical clustering and association rules – are applied on the simulation results to characterize important aspects for analysis.

Chapter 13, *A Geostatistically Based Probabilistic Risk Assessment Approach*, by Claudia Cherubini, poses the question of how to determine the levels of risk of contamination in environmental pollution risk assessment of zones, taking into consideration a high degree of variability and uncertainty that is inherent to the problem. The author uses an uncertainty modeling approach based on geostatistics for determining the parameters which enter as input to the probabilistic procedure of risk assessment. Although the focus of this work does not classify strictly as classical data mining but rather as statistical analysis, the discussion and approaches used provide insight and are relevant in any kind of analysis of spatial data.