

Foreword

A data warehouse is a centralized store of enterprise information implemented specifically for query, reporting and analysis purposes. Terrabytes of data are lying around throughout company networks that need to be integrated in data warehouses or data marts for analysis. One important problem is that the data lie in all sorts of heterogeneous systems, and therefore in all sorts of formats.

To integrate data in data warehouses, companies classically use extract, transform and load (ETL) software, which includes reading data from its source, cleaning it up and formatting it uniformly and then writing it to the warehouse to be exploited. Usually, data are gathered and transformed in simple relational tables organized in stars or snowflakes around fact tables.

With the explosion of multimedia data that are semi-structured or even unstructured, such as texts, images, sounds, videos, spatial information and XML documents, data warehouses need to be extended to support complex objects. The first part of this book is an interesting collection of papers dealing with the support of complex objects in a data warehouse. Seven papers investigate the following topics:

1. Spatial data modelling.
2. Design of XML document warehouses.
3. Warehousing active documents (i.e., XML documents including Web service calls).
4. Text warehousing.
5. Object-oriented source code integration and use for reengineering and reuse.
6. Managing metadata in decision environments.

7. A modelling framework for transaction datasets entailing a partitioned storage, efficient for data mining.

One major application of data warehouses is data mining. Data mining consists in extracting frequent patterns (tendencies, classes, rules, relationships, etc.) from databases. Storing complex objects in data warehouses requires exploring and modeling large data sets with varying composite structure or even without structure (e.g., free text). This issue of mining complex objects is addressed in the second part of this book. Six papers investigate the following topics:

8. Structural distance metrics to quantify the structural similarity between hierarchical structures, e.g., XML documents.
9. Approaches to evaluate and measure structural similarity of XML documents and schemas.
10. Storing, retrieving and manipulating patterns in an effective way using pattern models and pattern languages.
11. Multimedia database mining for visualization, using virtual reality.
12. Music data mining, including clustering, classification and pattern discovery in music.
13. Data mining for extracting meaningful biological information from genetic expressions.

Most papers include a survey of recent research and new proposals. All in all, this book is a significant collection of papers on new issues in warehousing complex data and using it for decision support. The papers have been carefully selected from a preliminary conference. Thirty three papers were submitted and 13 were selected. Altogether, the book constitutes a remarkable reference for researchers on new topics in online analysis and data mining.

Georges Gardarin

January 2006