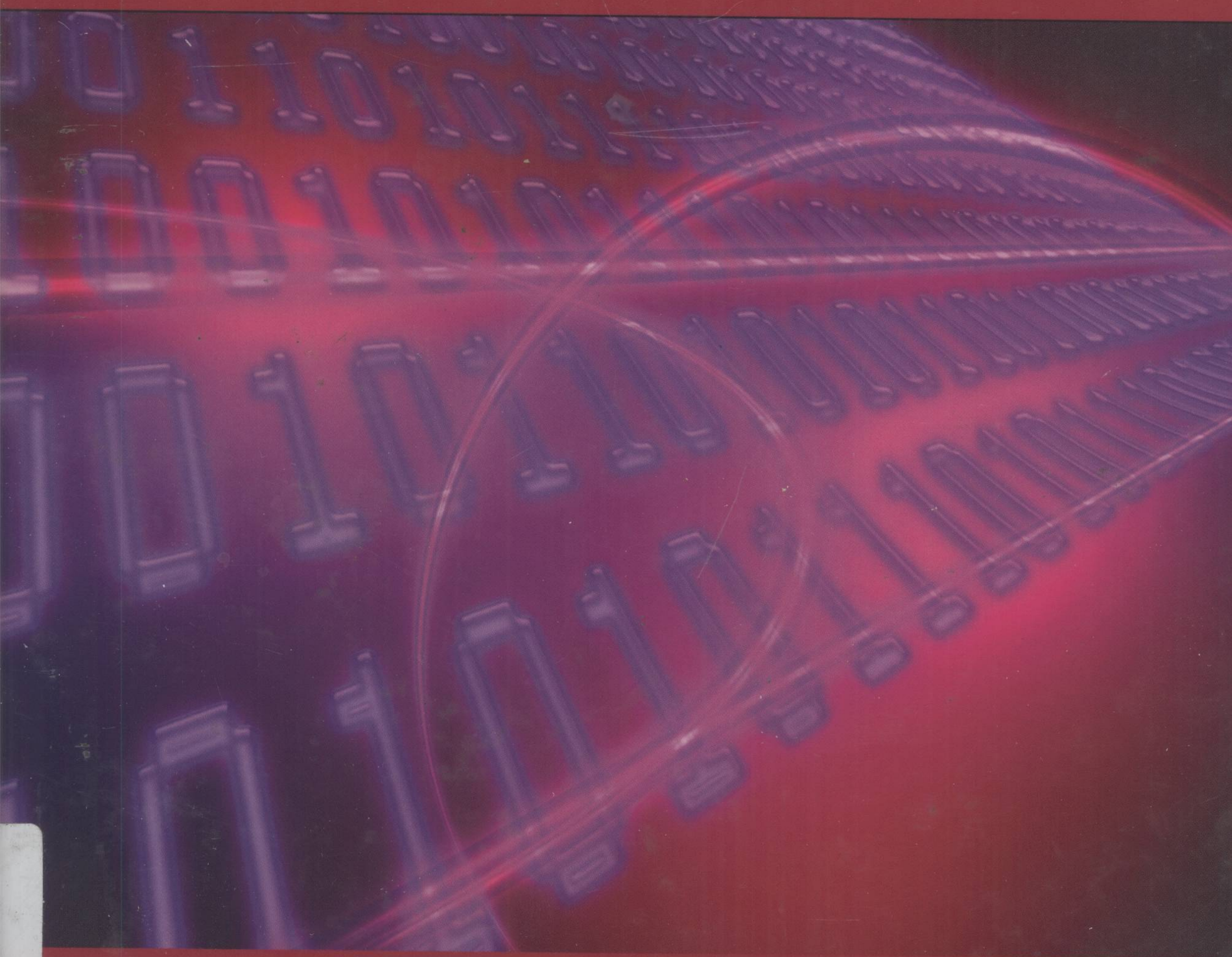
Data Mining and Knowledge Discovery Technologies



Data Mining and Knowledge Discovery Technologies

David Taniar
Monash University, Australia

IGIP
IGI PUBLISHING
Hershey • New York

Acquisition Editor:	Kristin Klinger
Senior Managing Editor:	Jennifer Neidig
Managing Editor:	Sara Reed
Development Editor:	Kristin Roth
Copy Editor:	Larissa Vinci
Typesetter:	Larissa Vinci
Cover Design:	Lisa Tosheff
Printed at:	Yurchak Printing Inc.

Published in the United States of America by
IGI Publishing (an imprint of IGI Global)
701 E. Chocolate Avenue
Hershey PA 17033
Tel: 717-533-8845
Fax: 717-533-8661
E-mail: cust@igi-global.com
Web site: http://www.igi-global.com

and in the United Kingdom by
IGI Publishing (an imprint of IGI Global)
3 Henrietta Street
Covent Garden
London WC2E 8LU
Tel: 44 20 7240 0856
Fax: 44 20 7379 0609
Web site: http://www.eurospanonline.com

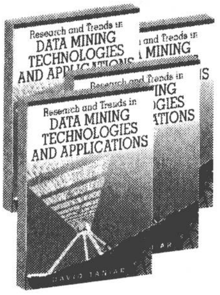
Product or company names used in this book are for identification purposes only. Inclusion of the names of the products or companies does not indicate a claim of ownership by IGI Global of the trademark or registered trademark.

Data mining and knowledge discovery technologies / David Taniar, editor.
p. cm. -- (Advances in data warehousing and mining, vol 2, 2007)
Summary: "This book presents researchers and practitioners in fields such as knowledge management, information science, Web engineering, and medical informatics, with comprehensive, innovative research on data mining methods, structures, tools, and methods, the knowledge discovery process, and data marts, among many other cutting-edge topics"--Provided by publisher.
Includes bibliographical references and index.
ISBN-13: 978-1-59904-960-1 (hardcover)
ISBN-13: 978-1-59904-961-8 (e-book)
1. Data mining. 2. Data marts. I. Taniar, David.
QA76.9.D343D3767 2007
005.74--dc22
2007037720

British Cataloguing in Publication Data
A Cataloguing in Publication record for this book is available from the British Library.

Data Mining and Knowledge Discovery Technologies is part of the IGI Global series named *Advances in Data Warehousing and Mining (ADWM)* (ISSN: 1935-2646).

All work contributed to this book is original material. The views expressed in this book are those of the authors, but not necessarily of the publisher.



Preface

This is the second volume of the Advances in Data Warehousing and Mining (ADWM) book series. ADWM publishes books in the areas of data warehousing and mining. The topic of this volume is data mining and knowledge discovery. This volume consists of 14 chapters in four section, contributed by authors and editorial board members from the International Journal of Data Warehousing and Mining, as well as invited authors who are experts in the data mining field.

Section I, *Association Rules,* consists of four chapters covering association rule techniques for multidimensional data, XML data, Web data, as well as rule interestingness measures.

Chapter I, *OLEMAR: An Online Environment for Mining Association Rules in Multidimensional Data,* by Riadh Ben Messaoud (University of Lyon 2), Sabine Loudcher Rabaséda (University of Lyon 2, France), Rokia Missaoui (University of Québec in Outaouais, Canada), and Omar Boussaid (University of Lyon 2, France), proposes to extend OLAP with data mining focusing on mining association rules in data cubes. OLEMAR (online environment for mining association rules) extracts associations from multidimensional data and allows extraction of inter-dimensional association rules, as well.

Chapter II, *Current Interestingness Measures for Association Rules: What do they Really Measure?*, by Yun Sing Koh (Auckland University of Technology, New Zealand), Richard O'Keefe (University of Otago, New Zealand), and Nathan Rountree (University of Otago, New Zealand), focuses on interestingness measurements for association rules. Rule interestingness measure is important as most of the association rule mining techniques, such as Apiori, commonly extract a very large number of rules, which might be difficult for decision makers to digest. It therefore makes sense to have these rules presented in a certain order or in groups or rules. This chapter studies the inter-relationship among variables in order to study the behaviour of the interestingness measures. It also introduces a classification of the current interestingness measures.

Chapter III, *Mining Association Rules from XML Data,* by Qin Ding and Gnanasekaran Sundarraj (The Pennsylvania State University at Harrisburg, USA), focuses on XML data.

XML data is growingly popular—used for data exchange as well as to represent semi-structured data. This chapter proposes a framework for association rule mining on XML data, and presents a Java-based implementation of the Apriori and FP-Growth algorithms for mining XML data.

Chapter IV, *A Lattice-Based Framework for Interactively and Incrementally Mining Web Traversal Patterns*, by Yue-Shi Lee and Show-Jane Yen (Ming Chuan University, Taiwan), concentrates on Web mining in order to improve Web services. It particularly focuses on Web traversal pattern mining, which discovers user access patterns from Web logs. This information is important, as it may be able to give Web users navigation suggestions. This chapter discusses efficient incremental and interactive mining algorithms to discover Web traversal patterns and make the mining results to satisfy the users' requirements.

Section II, *Clustering and Classification*, consists of four chapters covering clustering using genetic algorithm (GA) and rbf-Kernel, as well as classification of sequence data. This part also includes a chapter on the privacy issue.

Chapter V, *Determination of Optimal Clusters Using a Genetic Algorithm*, by Tushar, Shibendu Shekhar Roy, and Dilip Kumar Pratihar (IIT, Kharagpur), discusses the importance of clustering techniques. Besides association rules, clustering is an important data mining technique. A clustering method analyzes the pattern of a dataset and groups the data into several clusters based on the similarity among them. This chapter discusses clustering techniques using fuzzy c-means (*FCM*) and entropy-based fuzzy clustering (*EFC*) algorithms.

Chapter VI, *K-Means Clustering Adopting rbf-Kernel*, by ABM Shawkat Ali (Central Queensland University, Australia), focuses on the k-means clustering technique. This chapter presents an extension of the k-means algorithm by adding the radial basis function (rbf) kernel in order to achieve a better performance compared with the classical k-means algorithm.

Chapter VII, *Advances in Classification of Sequence Data*, by Pradeep Kumar (University of Hyderabad, Gachibowli, India), P. Radha Krishna (University of Hyderabad, Gachibowli, India), Raju S. Bapi (University of Hyderabad, Gachibowli, India), and T. M. Padmaja (University of Hyderabad, Gachibowli, India), focuses on sequence data. It reviews the state of the art for sequence data classification, including *kNN*, *SVM*, and Bayes classification. It describes the use of S^3M similarity metric. The chapter closes by pointing out various application areas of sequence data and describes open issues in sequence data classification.

Chapter VIII, *Using Cryptography for Privacy-Preserving Data Mining*, by Justin Zhan (Carnegie Mellon University, USA), focuses on privacy issues in *kNN* classification. Privacy concerns may prevent the parties from directly sharing the data and some types of information about the data. Therefore, the main issue is how multiple parties could share data in the collaborative data mining without breaching data privacy. The other issue is how to obtain accurate data mining results while preserving data privacy.

Section III on *Domain Driven and Model Free*, consists of two chapters covering domain driven and model free data mining.

Chapter IX, *Domain Driven Data Mining*, by Longbing Cao and Chengqi Zhang (University of Technology Sydney, Australia), proposes a practical data mining methodology called domain-driven data mining, whereby it meta-synthesizes quantitative intelligence and qualitative intelligence in mining complex applications. It targets actionable knowledge discovery in a constrained environment for satisfying user preference.

Chapter X, *Model Free Data Mining*, by Can Yang, Jun Meng, Shanan Zhu, and Mingwei Dai (Zhejiang University, Hangzhou, P. R. China and Xi'an Jiao Tong University, Xi'an, P. R. China), presents a model free data mining. This chapter shows the underlying relationship between sensitivity analysis and consistency analysis for input selection, and then derives an efficient model free method using common sense. It utilizes a fuzzy logic called fuzzy consistency analysis (*FCA*), which is a model free method and can be implemented efficiently as a classical model free method.

The final section, Section IV, *Issues and Applications*, consists of four chapters, discussing the minus sides of data mining, as well as presenting applications in bioinformatics and social sciences.

Chapter XI, *Minimizing the Minus Sides of Mining Data*, by John Wang (Montclair State University, USA), Xiaohua Hu (Drexel University, USA), and Dan Zhu (Iowa State University, USA), explores the effectiveness of data mining from a commercial perspective. It discusses several issues including the statistical issues, technical issues, and organizational issues.

Chapter XII, *Study of Protein-Protein Interactions from Multiple Data Sources,* by Tu Bao Ho, Thanh Phuong Nguyen, and Tuan Nam Tran (Japan Advanced Institute of Science and Technology, Japan), focuses on an application of data mining in the bioinformatics domain. This chapter gives a survey of computational methods for protein-protein interaction (PPI). It describes the use of inductive logical programming to learn prediction rules for protein-protein and domain-domain interactions.

Chapter XIII, *Data Mining in the Social Sciences and Iterative Attribute Elimination*, by Anthony Scime (SUNY Brockport, USA), Gregg R. Murray (SUNY Brockport, USA), Wan Huang (SUNY Brockport, USA), and Carol Brownstein-Evans (Nazareth College), presents an application in the social sciences domain. This domain is still underrepresented in the data mining area. With the large collection of social data, it gives potential opportunities to find society's pressing problems.

Finally, Chapter XIV, *A Machine Learning Approach for One-Stop Learning*, by Marco A. Alvarez and SeungJin Lim (Utah State University, USA), presents an application in the learning and education area. As the Web is nowadays an important source of learning, having an efficient tool and method for effective learning is critical. This chapter describes the use of *SVM*, AdaBoost, Naïve Bayes, and neural network in one-stop learning.

Overall, this volume covers important foundations to researches and applications in data mining, covering association rules, clustering, and classification, as well as new directions in domain driven and model free data mining. Issues and applications, particularly in bioinformatics, social and political sciences, and learning and education, show a full spectrum of the coverage of important and emerging topics in data mining.

David Taniar, Editor-in-Chief

Advances in Data Warehousing and Mining Series

November 2007

Section I

Association Rules

Data Mining and Knowledge Discovery Technologies

Table of Contents

Section II:
Clustering and Classification

Section III:
Domain Driven and Model Free

Section IV:
Issues and Applications

Chapter I

OLEMAR: An Online Environment for Mining Association Rules in Multidimensional Data

Riadh Ben Messaoud, University of Lyon 2, France
Sabine Loudcher Rabaséda, University of Lyon 2, France
Rokia Missaoui, University of Québec, Canada
Omar Boussaid, University of Lyon 2, France

Abstract

Data warehouses and OLAP (online analytical processing) provide tools to explore and navigate through data cubes in order to extract interesting information under different perspectives and levels of granularity. Nevertheless, OLAP techniques do not allow the identification of relationships, groupings, or exceptions that could hold in a data cube. To that end, we propose to enrich OLAP techniques with data mining facilities to benefit from the capabilities they offer. In this chapter, we propose an online environment for mining association rules in data cubes. Our environment called OLEMAR (online environment for mining association rules), is designed to extract associations from multidimensional data. It allows the extraction of inter-dimensional association rules from data cubes according to a sum-based aggregate measure, a more general indicator than aggregate values provided by the traditional COUNT measure. In our approach, OLAP users are able to drive a mining process guided by a meta-rule, which meets their analysis objectives. In *addition, the environment is based on a formalization, which exploits aggregate measures to revisit the definition of the support and the confidence of discovered rules. This formalization also helps evaluate the interestingness of association rules according to two additional quality measures: lift and loevinger. Furthermore, in order to focus on the discovered associations and validate them, we provide a visual representation based on the graphic semiology principles. Such a representation consists in a graphic encoding of frequent patterns and association rules in the same multidimensional space as the one associated with the mined data cube. We have developed our approach as a component in a general online analysis platform called Miningcubes according to an Apriori-like algorithm, which helps extract inter-dimensional association rules directly from materialized multidimensional structures of data. In order to illustrate the effectiveness and the efficiency of our proposal, we analyze a real-life case study about breast cancer data and conduct performance experimentation of the mining process.*

Introduction

Data warehousing and OLAP (online analytical processing) technologies have gained a widespread acceptance since the 90's as a support for decision-making. A data warehouse is a collection of subject-oriented, integrated, consolidated, time-varying, and non-volatile data (Kimball, 1996; Inmon, 1996). It is manipulated through OLAP tools, which offer visualization and navigation mechanisms of multidimensional data views commonly called *data cubes*.

A data cube is a multidimensional representation used to view data in a warehouse (Chaudhuri & Dayal, 1997). The data cube contains *facts* or *cells* that have *measures*, which are values based on a set of dimensions where each dimension usually consists of a set of categorical descriptors called *attributes* or *members*. Consider for example a *sales* application where the dimensions of interest may include, *costumer*, *product*, *location*, and *time*. If the measure of interest in this application is the *sales amount*, then an OLAP fact represents the sales measure corresponding to a single member in the considered dimensions. A dimension may be organized into a hierarchy. For instance, the location dimension may form the hierarchy *city* → *state* → *region*. Such dimension hierarchies allow different levels of granularity in the data warehouse. For example, a *region* corresponds to a high level of granularity whereas a *city* corresponds to a lower level. Classical aggregation in OLAP considers the process of summarizing data values by moving from a hierarchical level of a dimension to a higher one. Typically, additive data are suitable for simple computation according to aggregation functions (SUM, AVERAGE, MAX, MIN, and COUNT). For example, according to such a computation, a user may observe the sum of sales of products according to year and region.

Furthermore, with efficient techniques developed for computing data cubes, users have become widely able to explore multidimensional data. Nevertheless, the OLAP technology is quite limited to an exploratory task and does not provide automatic tools to identify and visualize patterns (e.g., clusters, associations) of huge multidimensional data.

In order to enhance its analysis capabilities, we propose to couple OLAP with data mining mechanisms. The two fields are complementary, and associating them can be a solution to cope with their respective limitations. OLAP technology has the ability to query and analyze multidimensional data through exploration, while data mining is known for its ability to discover knowledge from data. The general issue of coupling database systems with data mining was already discussed and motivated by Imieliński and Mannila (1996). The authors state that data mining leads to new challenges in the database area, and to a second generation of database systems for managing KDD (knowledge discovery in databases) applications just as classical ones manage business ones. More generally, the association of OLAP and data mining allows elaborated analysis tasks exceeding the simple exploration of data. Our idea is to exploit the benefits of OLAP and data mining techniques and to integrate them in the same analysis framework. In spite of the fact that both OLAP and data mining were considered two separate fields for a while, several recent studies showed the benefits of coupling them.

In our previous studies, we have shown the potential of coupling OLAP and data mining techniques through two main approaches. Our first approach deals with the reorganization of data cubes for a better representation and exploration of multidimensional data (Ben Messaoud, Boussaid, & Loudcher, 2006a). The approach is based on multiple correspondence analysis (MCA), which allows the construction of new arrangements of modalities in each dimension of a data cube. Such a reorganization aims at bringing together cells in a reduced part of the multidimensional space, and hence giving a better view of the cube. Our second approach constructs a new OLAP operator for data clustering called *OpAC* (Ben Messaoud, Boussaid, & Loudcher, 2006b), which is based on the agglomerative hierarchical clustering (AHC).

In this chapter, we present a third approach which also follows the general issue of coupling OLAP with data mining techniques but concerns the mining of association rules in multidimensional data. In Ben Messaoud, Loudcher, Boussaid, and Missaoui (2006), we have proposed a guided-mining process of association rules in data cubes. Here, we enrich this proposal and establish a complete online environment for mining association rules (*OLEMAR*). In fact, it consists of a mining and visualization package for the extraction and the representation of associations from data cubes. Traditionally, with OLAP analysis, we used to observe summarized facts by aggregating their measures according to groups of descriptors (members) from analysis dimensions. Here, with *OLEMAR*, we propose to use association rules in order to better understand these summarized facts according to their descriptors. For instance, we can note from a given data cube that sales of *sleeping bags* are particularly high in a given city. Current OLAP tools do not provide explanations of such particular fact. Users are generally supposed to explore the data cube according to its dimensions in order to manually find an explanation for a given phenomenon. For instance, one possible interpretation of the previous example consists in associating sales of *sleeping bags* with the *summer season* and *young tourist costumers*.

In the recent years, many studies addressed the issue of performing data mining tasks on data warehouses. Some of them were specifically interested in mining patterns and association rules in data cubes. For instance, Kamber, Han, and Chiang (1997) state that it is important to explore data cubes by using association rule algorithms. Further, Imieliński, Khachiyan, and Abdulghani (2002) believe that OLAP is closely interlinked with association rules and shares with them the goal of finding patterns in the data. Goil and Choudhary (1998) argue that automated techniques of data mining can make OLAP more useful and easier to apply in the overall scheme of decision support systems. Moreover, cell frequencies can facilitate the computation of the support and the confidence, while dimension hierarchies can be used to generate multilevel association rules.

OLEMAR is mainly based on a mining process, which explains possible relationships of data by extracting *inter-dimensional* association rules from data cubes (i.e., rules mined from multiple dimensions without repetition of predicates in each dimension). This process is guided by the notion of *inter-dimensional meta-rule,* which is designed by users according to their analysis needs. Therefore, the search of association rules can focus on particular regions of the mined cube in order to meet specific analysis objectives. Traditionally, the COUNT measure corresponds to the frequency of facts. Nevertheless, in an analysis process, users are usually interested in observing multidimensional data and their associations according to measures more elaborated than simple frequencies. In our approach, we propose a redefinition of the support and the confidence to evaluate the interestingness of mined association rules when *SUM-based* measures are used. Therefore, the support and the confidence according to the COUNT measure become particular cases of our general definition. In addition to support and confidence, we use two other descriptive criteria (*lift* and *loevinger*) in order to evaluate the interestingness of mined associations. These criteria are also computed for *sum-based aggregate measures* in the data cube and reflect interestingness of associations in a more relevant way than what is offered by support and confidence.

The mining algorithm works in a *bottom-up* manner and is an adaptation of the *Apriori* algorithm (Agrawal, Imieliński, & Swami, 1993) to multidimensional data. It is also guided by user's needs expressed through the meta-rule, takes into account a user selected measure in the computation of the support and the confidence, and provides further evaluation of extracted association rules by using *lift* and *loevinger* criteria.

In addition to the mining process, the environment also integrates a visual tool, which aims at representing the mined frequent patterns and the extracted association rules according to an appropriate graphical encoding based on the *graphic semiology* principles of Bertin (1981). The peculiarity of our visualization component lies on the fact that association rules are represented in a multidimensional space in a similar way as facts (cells).

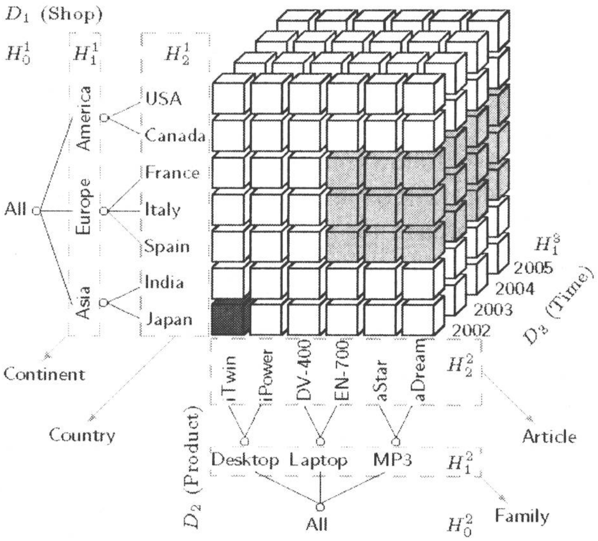
This chapter is organized as follows. In the second section, we define the formal background and notions that will be used in the sequel. The third section presents the key concepts of our approach for mining inter-dimensional association rules: the concept of inter-dimensional meta-rule; the general computation of support and confidence based on OLAP measures; and criteria for the advanced evaluation of mined association rules. The fourth section deals with the visualization of the mined inter-dimensional association rules while the fifth section provides the implementation of the online mining environment and describes our algorithm for mining inter-dimensional association rules. In the sixth section, we use a case study about mammographies to illustrate our findings while the seventh section concerns the experimental analysis of the developed algorithm. In the eighth section, we present a state of the art about mining association rules in multidimensional data. We also provide a comparative study of existing work and our own proposal. Finally, we conclude this chapter and address future research directions.

Formal Background and Notations

In this section, we define preliminary formal concepts and notations we will use to describe our mining process. Let C be a data cube with a non empty set of d dimensions $\mathbf{D} = \{D_1, \ldots, D_i, \ldots, D_d\}$ and a non empty set of measures $\mathbf{M}$. We consider the following notations:

- Each dimension $D_i \in \mathbf{D}$ has a non empty set of hierarchical levels. C;
- H_j^i is the j^{th} $(j \geq 0)$ level hierarchical level in D_i. The coarse level of D_i, denoted H_0^i, corresponds to its total aggregation level *All*. For example, in Figure 1, dimension *Shop* (D_1) has three levels: *All*, *Continent*, and *Country*. The *All* level is denoted H_0^1, the *Continent* level is denoted H_1^1, and the *Country* level is denoted H_2^1;

Figure 1. Example of sales data cube



- $\mathbf{H}_i$ is the set of hierarchical levels of dimension D_i, where each level $H_j^i \in \mathbf{H}_i$ consists of a non empty set of members denoted A_{ij}. For example, in Figure 1, the set of hierarchical levels of D_2 is $\mathbf{H}_2 = \{H_0^2, H_1^2, H_2^2\} = \{$*All, Family, Article*$\}$, and the set of members of the *Article* level of D_2 is $A_{22}=\{$*iTwin, iPower, DV-400, EN-700, aStar, aDream*$\}$.

Definition 1. (Sub-cube)

Let $\mathbf{D} \subseteq \mathbf{D}$ be a non empty set of p dimensions $\{D_1, \ldots, D_p\}$ from the data cube C ($p \leq d$). The p-tuple $(\Theta_1, \ldots, \Theta_p)$ is called a sub-cube on C according to $\mathbf{D}$ iff $\forall i \in \{1, \ldots, p\}$, $\Theta_i \neq \emptyset$ and there exists a unique j such that $\Theta_i \subseteq A_{ij}$.

As previously defined, a sub-cube according to a set of dimensions $\mathbf{D}$ corresponds to a portion from the initial data cube C. It consists in setting for each dimension from $\mathbf{D}$ a non-empty subset of member values from a single hierarchical level of that dimension. For example, consider $\mathbf{D} = \{D_1, D_2\}$ a subset of dimensions from the cube of Figure 1. $(\Theta_1, \Theta_2) = ($*Europe*, $\{$*EN-700, aStar, aDream*$\})$ is therefore a pos-