Preface

During the last decade, the French-speaking scientific community developed a very strong research activity in the field of Knowledge Discovery and Management (KDM or EGC for "Extraction et Gestion des Connaissances" in French), which is concerned with, among others, Data Mining, Knowledge Discovery, Business Intelligence, Knowledge Engineering and Semantic Web.

This emerging research area has also been strongly stimulated by the rapid growth of information systems and the web semantic issues. The success of the first two French-speaking EGC Conferences, held in Nantes in 2001 and in Montpellier in 2002, resulted naturally in 2003 in the foundation of the International French speaking EGC Association[1]. Since, the Association yearly organizes conferences and workshops with the aim of promoting exchanges between researchers and companies concerned with KDM and its application in business, administration, industry or public organizations.

The recent and novel research contributions collected in this book are extended and reworked versions of a selection of the best papers that were originally presented in French at the EGC 2010 Conference held in Hammamet Tunisia, on January 2010. The 12 best papers that have been selected for this book are issued from the 32 papers accepted in long format with a 23% acceptance ratio among the 139 papers submitted to the EGC2010 conference[2].

Structure of the book

The volume is organized in three parts.

[1] Association "Extraction et Gestion des Connaissances" (EGC), see `http://www.egc.asso.fr`

[2] For further details about EGC2010 conference see "10^{th} International French-Speaking Conference on Knowledge Discovery and Management (EGC2010): conference report" by Sadok Ben Yahia and Jean-Marc Petit, in ACM SIGKDD Explorations, volume 12, issue 1

Part I – Various aspects of Data Cube and Ontology-based representations

In the first chapter, entitled *Constrained Closed and Quotient Cubes*, Rosine Cicchetti, Lotfi Lakhal and Sébastien Nedjar investigate reduced representations for the Constrained Cube. They propose two representations which discard redundancies, are information lossless and avoid to compute the whole data cubes: the Constrained Closed and Quotient Cube. Provided with the former, the decision maker can perform OLAP classification and querying. The latter adds navigation within the cube to the previous capabilities. Hence according to her/his needs, the user can choose the more suitable representation.

In the second chapter, *A New Concise and Exact Representation of Data Cubes*, Hanen Brahmi, Tarek Hamrouni, Riadh Ben Messaoud and Sadok Ben Yahia introduce a new concise and exact representation called closed non derivable data cubes (CND-Cube). It is based on the concept of non derivable minimal generators. They also propose a novel algorithm dedicated to the mining of CND-Cube from multidimensional databases and investigate in detail the theoretical foundations of this concise representation. Moreover, they discuss Rollup/Drilldown semantics over the CND-Cube and validate their approach on a set of real and benchmark datasets. Their experiment results show the effectiveness of the approach in comparison with those fitting in the same trend.

Michel Buffa and Catherine Faron-Zucker are interested in *Ontology-Based Access Rights Management*. They propose an approach to manage access rights in a content management systems which relies on semantic web models and technologies. They developed the AMO ontology which consists in 1) a set of classes and properties dedicated to the annotation of resources whose access should be controlled and in 2) a base of inference rules modelling the access management strategy to carry out. When applied to the annotations of the resources whose access should be controlled, these rules enable to manage access according to a given strategy. This modelisation is flexible, extendable and ensures the adaptability of the AMO ontology to any access management strategy.

In the last chapter of this first part, Souhir Gahbiche, Nathalie Pernelle and Fatiha Saïs addresse the problem of explanation of data reconciliation decisions that are obtained by automatic numerical informed by ontology knowledge data reconciliation methods. These methods may give rise to decision errors and to approximated results. They have developed an explanation model based on Coloured Petri Nets formalism. This model allows to generate a graphical and readable explanation, which takes into account all the semantic knowledge that are involved in the similarity computation of a reference pair.

Part II – Efficient Pattern Mining issues

The first of chapter of this second part by Mehdi Khiari and Patrice Boizumault and Bruno Crémilleux proposes to *combine Constraint Programming and Constraint-based Mining*. By investigating the relationship between constraint-based mining and constraint satisfaction problems, they propose a generic approach to model and mine queries involving several local patterns (n-ary patterns). The resulting framework for pattern set mining is very flexible and allows the user to easily express in a declarative way a wide range of problems for a wide range of data mining tasks.

The next chapter written by Marc Boullé is concerned with *Simultaneous Partitioning of Input and Class Variables for Supervised Classification Problems with Many Classes*. It extends discretization and value grouping methods, based on the partitioning of both the input and class variables. The best joint partitioning is searched by maximizing a Bayesian model selection criterion. This preprocessing is exploited as a preparation for the naive Bayes classifier, and demonstrates high performance in problems with hundreds of classes.

Lionel Martin and his co-authors describe an *Interactive and progressive constraint definition for dimensionality reduction and visualization*. They propose a tool for semi-supervised, constraint-based projection of data. Starting from an initial linear projection, a user can specify constraints, in order to move away or closer of some pairs of objects. Based on the Uzawa algorithm, a new projection is computed that takes these constraints into account. The user can iteratively add new constraints. The implementation is based on Explorer3D, the 3D projection tool developed at the LIFO (Computer Science Laboratory of Orleans, France).

The last chapter of this part by Anne Laurent and his co-authors deals with *Efficient parallel mining of gradual patterns on multicore processors*. Mining gradual patterns plays a crucial role in many real world applications where huge volumes of complex numerical data must be handled, *e.g.*, biological databases, survey databases, data streams or sensor readings. Gradual patterns highlight complex order correlations of the form “The more/less X, the more/less Y”. They present an efficient parallel algorithm to mine gradual patterns on multicore processors. They experimentally show that this algorithm significantly improves the state of the art and scales very well with the number of cores available.

Part III – Data Preprocessing and Information Retrieval

In the first one, entitled *Analyzing Old Documents Using a Complex Approach: application to lettrines indexing*, by Mickael Coustaty and his co-authors proposes a methodology based on a complex approach to analyze and to characterize graphical images extracted from old documents. Based on a comparison between historians and computer science approaches, the authors propose a novel method to describe images using a complex approach to globally describing an image with respect to its structure, sense, and elements.

In *Identifying relevant features of images from their 2-D topology*, Marc Joliveau introduces a new method for the visual perception problem that focuses on the intrinsic two dimensional topology of images to extract their principal features. Validations on three different datasets show that the few features extracted by the method provide enough intelligible information to efficiently and fastly identify similarities between the images of a database.

The next chapter by Radja Messai *et al.* entitled *Analyzing Health Consumer Terminology for Query Reformulation Tasks* describes the analysis and characterisation of the health consumer terminology in the breast cancer field. The results have been used in a pilot study to enhance the reformulation of health consumers' queries. The results have showed significant differences between the health consumer terminology and the professional one. Such studies are important to provide health services more adapted to the language and the level of knowledge of health consumers.

The last chapter, proposed by Patrick Bosc *et al.*, addresses the plethoric answer problem that often arises when end-users have an approximate idea of how to formulate a query to retrieve what they are looking for from large-scale databases. A possible approach to reduce the set of retrieved items and to make it more manageable is to constrain the initial query with additional predicates. The approach presented in this paper relies on the identication of correlation links between predefined predicates related to attributes of the relation of interest. Thus, the initial query is strengthened by additional predicates that are semantically close to the user-specified ones.

Acknowledgments

The editors would like to thank the chapter authors for their insights and contributions to this book.

The editors would also like to acknowledge the members of the review committee and the associated referees for their involvement in the review process of the book. Their in depth reviewing, criticisms and constructive remarks significantly contributed to the high quality of the retained papers.

A special thank goes to Bruno Pinaud who has efficiently composed and laid out the manuscript.

Finally, we thank Springer and the publishing team, and especially T. Ditzinger and J. Kacprzyk, for their confidence in our project.

Nantes, Geneva, Lyon
September 2011

Fabrice Guillet
Gilbert Ritschard
Djamel A. Zighed

Review Committee

All published chapters have been reviewed by at least 2 referees.

- Tomas Aluja (UPC, Barcelona, Spain)
- Nadir Belkhiter (Univ. Laval, Québec, Canada)
- Sadok Ben Yahia (Univ. Tunis, Tunisia)
- Younès Bennani (Univ. Paris 13, France)
- Omar Boussaid (Univ. Lyon 2, France)
- Paula Brito (Univ. of Porto, Portugal)
- Francisco de A. T. De Carvalho (Univ. Federal de Pernambuco, Brazil)
- Gilles Falquet (Univ. of Geneva, Switzerland)
- Jean-Gabriel Ganascia (Univ. Paris 6, France)
- Pierre Gancarski (Univ. of Strasbourg, France)
- Howard Hamilton (Univ. of Regina, Canada)
- Robert Hilderman (Univ. of Regina, Canada)
- Petra Kraemer (Univ. La Rochelle, France)
- Philippe Lenca (Telecom Bretagne, Brest, France)
- Philippe Leray (Univ. of Nantes, France)
- Stan Matwin (Univ. of Ottawa, Canada)
- Monique Noirhomme (FUNDP, Namur, Belgium)
- Matthieu Perreira Da Silva (Univ. La Rochelle, France)
- Vincent Pisetta (Univ. Lyon 2, France)
- Pascal Poncelet (LIRMM, Univ. of Montpellier, France)
- Zbigniew Ras (Univ. of North Carolina, USA)
- Jan Rauch (University of Prague, Czech Republic)
- Chiara Renso (KDDLAB - ISTI CNR, Italy)
- Lorenza Saitta (Univ. di Torino, Italy)
- Ansaf Salleb-Aouissi (Columbia Univ., New York, USA)
- Florence Sédes (Univ. of Toulouse 3, France)
- Arno Siebes (Univ. Utrecht, The Netherlands)
- Dan Simovici (Univ. of Massachusetts Boston, USA)
- Stefan Trausan-Matu (Univ. of Bucharest, Romania)
- Rosanna Verde (Second Univ. of Naples, Italy)
- Christel Vrain (Univ. of Orléans, France)
- Jef Wijsen (Univ. of Mons-Hainaut, Belgium)
- Chongsheng Zhang (Univ. of Nice, France)

Associated Reviewers

Yassine Benabbas,
Patrick Bosc,
Marc Boullé,
Hanen Brahmi,
Guillaume Cleuziou,
Mickaël Coustaty,
Bruno Cremilleux,
Catherine Faron Zucker,
Manfredotti,
Anne Laurent,
Lionel Martin,
Radja Messai,
Jonathan Weber

Manuscript Coordinator

Bruno Pinaud (Univ. of Bordeaux I, France)

Contents

Part III Data Preprocessing and Information Retrieval