Revue des Nouvelles Technologies de l'Information
Sous la direction de Djamel A. Zighed et Gilles Venturini

RNTI E.29

Symbolic data analysis and Visualisation: Special Issue in honor of Monique Noirhomme-Fraiture

Rédacteurs invités/Guest Editors: Paula Brito (University of Porto, Portugal),
Gilles Venturini (University François Rabelais of Tours, France)

LE MOT DES DIRECTEURS DE LA COLLECTION RNTI
FOREWORD OF THE COLLECTION'S DIRECTORS

Très chers lecteurs,

La revue RNTI (www.editions-rnti.fr) publie environ 2000 pages par an en appliquant des critères de qualité bien définis. Les thèmes de la revue sont liés à l'Extraction de connaissances à partir des données, à la Fouille de données et à la Gestion des connaissances, mais également à d'autres domaines de l'Informatique. Nous vous invitons à proposer des projets éditoriaux rentrant dans la politique éditoriale de RNTI et dont les principes assez simples font la distinction entre deux sortes de publications :

- des numéros thématiques faisant l'objet d'un appel à communication, avec un ou plusieurs rédacteurs invités,
- des actes de conférences sélectives garantissant une haute qualité des articles.

Nous remercions chaleureusement les rédacteurs invités de ce numéro pour la confiance qu'ils portent dans RNTI. Pour tout renseignement, nous vous invitons à consulter notre site Web et à nous contacter.

Djamel A. Zighed and Gilles Venturini.

Dear readers,

The RNTI journal (New IT journal, www.editions-rnti.fr) publishes about 2000 pages per year by applying well-defined quality criteria. The themes of the journal are related to Knowledge Discovery from Databases, to Data Mining, to Knowledge Management, but also to other areas of IT or Computer Science. We invite you to submit editorial projects falling within the editorial policy of RNTI, which distinguishes between two kinds of publications:

- thematic issues with a call for papers, and with one or more guest editors,
- selective conference proceedings providing high quality papers.

We would like to thank the guest editors of this issue for their confidence in RNTI. For more information, please visit our website and contact us.

PRÉFACE/FOREWORD

It is a great pleasure for us to publish this special issue of RNTI in honour of Monique Noirhomme-Fraiture. This issue presents selected papers on two domains where Monique has concentrated her research these last years: Data Visualization and Symbolic Data Analysis.

In Data Visualization, one of the mains issue is to augment standard data analysis and data mining methods with visual and interactive approaches, so as to make such methods closer to the user's needs. Researchers who consider that the user is at the centre of the Data Mining process are indeed concerned with Data Visualization, and more generally, with Human-Computer Interaction. This user-centred point of view is often mandatory because in Knowledge Discovery only the domain expert can evaluate and validate the results. Visual approaches can communicate a high amount of information about the data and the discovered knowledge. Furthermore, with appropriate interactions, the expert can participate in the discovery process, and can drive automatic tools towards better results.

One of the papers in this special issue deals with the use of data visualization to improve community detection in graphs and illustrate those principles. Other papers in this issue are also using visual and interactive methods.

Symbolic Data Analysis (SDA) is a relatively recent field of research focusing on the representation and analysis of data presenting internal variability. In Data Mining and classical Statistics the entities to be analysed usually present one single value for each variable, that is however no longer the case when we analyse groups gathered on the basis of some given criteria and not single elements: the variability inherent to each group for the different variables should be taken into account. When analysing concepts, such as botanic species, disease descriptions, car models, etc., data also comprehends intrinsic variability that must be considered. To this purpose, new variable types have been introduced, whose realizations are not single real values or categories, but sets, intervals, or, more generally, distributions over a given domain. In the context of SDA, methods for the (multivariate) analysis of such "symbolic" data have been developed, where the variability expressed in the data representation is taken into account.

Seven of the papers in this Special Issue come within the framework of SDA, addressing distinct problems, from database generalization to metadata representation, descriptive order statistics and corresponding visualization and clustering. Applications in different domains show that the consideration of symbolic data proves to be an adequate and useful approach in distinct situations. Most of these SDA papers address visualization issues, thereby providing contributions for the visual representation of symbolic data and/or the results of symbolic data analysis.

Below we provide a short description of the accepted articles:

- The paper "Visualization-based communities discovering in commuting networks : a case study" by François Queyroi and Yves Chiricota proposes a procedure to identify hierarchical partitions of cities in a given territory that captures commuters flows density. The method is applied on a network which represents commuting in France. The authors' approach is based on a common technique improved by visual tools: highlight dense areas using a strength metric and extract clusters at different levels using the variation of a quality measure function.

- The paper "Hierarchical clustering of modal ordinal symbolic data objects" by Carmen Bravo and José M. García-Santesmases addresses the problem of evaluating the dispersion of a set of objects described by ordinal modal symbolic data, in a clustering context. A consensus measure for objects and for sets of objects described by modal ordinal data is defined, and a variability measure for sets of subsets of objects based in the consensus measure of their members is proposed.This then leads to a dissimilarity measure between objects and between set of objects. An ascending hierarchical clustering algorithm is presented where the criterion to be minimized in each step is based on the decrease of the consensus variability. The method is illustrated with modal ordinal data obtained from 34 teachers evaluation. A visual representation of the teachersŠ modal ordinal symbolic data is provided, and also applied to the cluster centroids.

- The paper "Order statistics for histogram data and a blox plot visualization tool" by Rosanna Verde, Antonio Balzanella and Antonio Irpino proposes new descriptive statistics for symbolic histogram data. The authors define median and quartiles of a histogram variable using the quantile functions associated with the corresponding empirical distribution functions of the observed histograms. The definition of an order relationship between quantile functions is based on the ℓ^p Wasserstein distance. The classic box-plot representation is then extended to quantile functions. New measures of variability and skewness for a histogram variable are defined, associated with this representation. An application on real data regarding climatic information recorded at 60 meteorological Chinese stations illustrates the proposed measures and the new box-plot visualization tool.

- In the paper "Classical and Symbolic metadata setting for biological datasets", Haralambos Papageorgiou and Maria Vardaki use symbolic data analysis to describe the management process of biological datasets produced by multi-source clinical studies. This naturally leads to more complex data types and tables, so that the metadata under consideration hold information both on the classical (original) and on the symbolic data. For this purpose, the authors introduce an abstract object-oriented metadata model designed in Unified Modeling Language that can hold metainformation both for the classical (original) clinical data and the metadata for the symbolic data setting. A number of transformations are discussed both for classical and symbolic classes of the model, illustrating how the applied transformations on symbolic data depend on the related classical data setting.

- The paper "Generalization Method when Manipulating Relational Databases" by V. Cariou and L. Billard addresses the problem of aggregating large datasets in the form of interval data. The authors propose algorithms to build intervals which are typically homogeneous, and to test this homogeneity. Aggregation of data into intervals is made by combining a generalization operator and an associated reduction algorithm, so as to obtain intervals which closely reflect the original dataset. A test is also proposed for the hypothesis that observations across the resulting intervals are mixtures of uniform distributions rather than a single distribution, which include consideration about outlier observations. The proposed methods are illustrated on two synthetic datasets.

- The paper "Analyzing European Social Survey data using symbolic data methods and Syrokko software" by Filipe Afonso and Seppo Laaksonen presents an application of Symbolic Data Analysis (SDA) with SYR software to the fifth round of the European Social Survey (ESS) carried out among European inhabitants. The study focuses on the 52 European countries by age groups, which are described by interval-valued and discrete-distribution-valued (bar-chart) variables. The paper presents some functionalities of the SYR software, and uses it to analyse the symbolic data. Interval and bar-chart descriptions allow comparing the country × age groups for individual variables. Then Principal Component Analysis, followed by K-means clustering on the factorial coordinates provide a multivariate analysis of the data. The cluster prototypes are also represented within the same language. Biplots in the first factorial plane provide a graphical representation, to which cluster descriptions may be associated.

- In the paper "Analyse symbolique de sourires de personnages virtuels" by Magalie Ochs, Edwin Diday and Filipe Afonso, the authors analyse a database of smiles of a virtual character, directly created by users. The data array is composed by 5517 smile descriptions, of which 2057 are amused smiles, 1675 are polite smiles and the remaining 1785 are embarrassed smiles, each described by eight variables. Each type of smile is then associated with a symbolic description in terms of discrete frequency distributions. To identify the morphological and dynamic characteristics of the different types of smile, symbolic descriptive and supervised analysis have been applied. A non-supervised analysis then allowed identifying other types of smiles corresponding to particular combinations. Principal Component Analysis of the partitions' prototypes descriptions provides a visual representation of the different classes obtained. A further clustering of the classes' prototypes identifies five clusters, corresponding to amused smiles, polite smiles, embarrassed smiles, combination of embarrassed-polite and combination of amused-polite.

- The paper "Cartes auto-organisatrices pour la classification des données de type intervalle en se basant sur la distance city-block", by Chantal Hajjar and Hani Hamdan proposes an algorithm to train self-organizing maps in batch mode, in order to classify interval data. The process is based on the optimization of an adequacy criterion based on the city-block distance. The choice of the mapŠs parameters and data normalization are discussed. The method is tested and compared to other clustering methods for interval data using a data set of 33 car models and a data set of minimal and maximal temperatures registered in 37 cities around the world during the 12 months of the year.

The papers in this Special Issue constitute a selection of recent research and applications in Data Visualization and Symbolic Data Analysis. The variety of data structures, problems addressed and methodologies developed cleary show that these are dynamical fields of research, where we will certainly witness much and exciting developments in the times ahead. We hope the reader finds these papers interesting and feels motivated by these challenging areas.

The Editors gratefully acknowledge the invaluable assistance of the experts and colleagues in the process of reviewing the manuscripts that were submitted for this Special Issue.

Paula Brito
University of Porto
Portugal

Gilles Venturini
University François Rabelais of Tours
France

TABLE DES MATIÈRES/TABLE OF CONTENTS