Editorial

Alfredo Cuzzocrea

ICAR-CNR and University of Calabria,
Via P. Bucci, 41C, Rende,
87036 Cosenza, Italy
E-mail: cuzzocrea@si.deis.unical.it

The problem of mining, modelling and managing complex data arising in data- and knowledge-intensive next-generation information systems plays a critical role in actual *databases and information systems* research. Complex data are characterised by several and heterogeneous formats: transactional data, multidimensional data, hierarchical data, semi-structured data, biological data, ontological data, streaming data, trajectory data, moving object data, and so forth. Traditional models, techniques and algorithms for representing, querying and mining conventional datasets (e.g., relational ones) are widely recognised as inadequate to cope with challenges posed by processing complex data. Therefore, designing and developing suitable models, techniques and algorithms for processing complex data has become an emerging challenge in actual databases and information systems research. This trend has also been dictated and stirred-up by the recent explosion of complex *intelligent information systems*, which are becoming more and more popular in a wide range of application scenarios ranging from *complex database*/*data-mining systems* to *data-warehouse*/*business-intelligence systems*, from *grid computing* to *peer-to-peer computing*, from *biological data management tools* to *ontology-based complex tools*, and so forth.

This special issue on 'Advances in mining, modelling and managing complex data' of the *International Journal of Data Mining, Modelling and Management* focuses on latest research results and open research challenges on the problem of effectively and efficiently mining, modelling and managing complex data, according to principles and guidelines provided above.

With the aim of adequately fulfilling both theoretical and practical issues deriving from mining, modelling and managing complex data, this special issue contains five papers, which have gone through two-rigorous review rounds before being accepted for the final inclusion. Some of the contributions of this special issue have been invited for submission as best papers from the session 'Advanced Knowledge-Based Systems' of the 12th International Conference on Knowledge-based and Intelligent Information and Engineering Systems (KES 2008), held in Zagreb, Croatia, during September 3–5, 2008, leaded by the editor.

The first paper, titled 'Parallel hierarchical clustering using weighted confidence affinity', by Baoying Wang, Imad Rahal and Aijuan Dong, proposes a *parallelised hierarchical clustering approach for categorical data*, called *PH-clustering*, using vertical data structures. Main motivations of authors start from recognising that many attempts for clustering categorical data such as market basket dataset exist, but most of categorical clustering approaches belong to partitional clustering which requires at least one input parameter (e.g., the minimum intra-cluster similarity or the desired number of

Copyright © 2011 Inderscience Enterprises Ltd.

clusters). This can turn off in a serious limitation of the quality of final clusters. In order to minimise the impact of low support items, authors devise a *weighted confidence* (WC) affinity function to compute the similarity between clusters. Based on the analysis of the major clustering steps, authors adopt a *partial local* and *partial global* approach to reduce computation time as well as to keep network communication at minimum. Load balance issues are addressed especially during the data partitioning phase. Experimental results on standardised market basket data show that the proposed WC affinity measure is more accurate than other contemporary affinity measures in the literature and that the parallel clustering approach provides magnitudes of time improvements over sequential clustering especially over larger data sizes. Results also indicate that the number of items/attributes in the dataset has a more drastic impact on performance than the number of transactions/tuples.

The second paper, titled 'A pattern matching approach for clustering gene expression data', by Rosy Das, Jugal Kalita and Dhruba K. Bhattacharyya, focuses the attention on the problem of *identifying groups of genes with similar expression time courses during the analysis of gene expression time series data*. A *regulation-based clustering approach*, named as *PatternClus*, for clustering gene expression data is proposed to this aim. The proposed method is also able of identifying sub-clusters based on an order-preserving ranking approach. *PatternClus* is experimentally assessed on real life datasets, and it is established to perform satisfactorily. *PatternClus* is also compared to some well-known clustering algorithms (*k-means* and *hierarchical algorithms*). This experimentation demonstrates that *PatternClus* provides better results in terms of z-score measure of cluster validation. An incremental version of *PatternClus* is also presented, which helps in identifying clusters incrementally when the underlying database is continuously increasing.

In the third paper, titled 'Efficient evaluation of partially-dimensional range queries in large OLAP datasets', by Yaokai Feng, Kunihiko Kaneko and Akifumi Makinouchi, authors address the increasing requirement for *processing multidimensional queries on OLAP datasets* that has characterised the database community during the last decades, with particular emphasis on evaluating *range queries* on large OLAP datasets. As authors correctly observe, *multidimensional indices* are helpful to improve the performance of such queries. However, much information irrelevant to queries has to be read from disk if the existing multidimensional indices are used with OLAP data. This greatly degrades the effectiveness and the efficiency of the whole query evaluation task. In light of this, authors propose a novel *index structure for multidimensional data*, called *AR*-tree*, inspired from the well-known *R*-tree*, for *efficiently evaluating partially-dimensional* (PD) *range queries on large OLAP datasets*. PD range queries are range queries such that query conditions probably are only with partial dimensions (not all) of the whole index space. As authors correctly state, PD range queries are popular operations in modern OLAP applications, hence, optimising such queries plays a critical role. In order to speed-up query execution time, *AR*-tree* allows us to *counter* the actual queries on target OLAP data, hence, providing significant complexity reduction. Results of both mathematical analysis and a comprehensive campaign of experiments on different classes of OLAP datasets indicate that *AR*-tree* can clearly improve the performance of PD range queries, especially when large OLAP dataset instances are considered.

The fourth paper, titled 'Privacy preserving association rules mining on distributed homogenous databases', by Mahmoud Hussein, Ashraf El-Sisi and Nabil Ismail, proposes a novel version of classical *privacy preserving association rule mining*

algorithms on distributed homogenous databases, whose main goal consists in overcoming state-of-the-art approaches for what regards both performance and accuracy of final results. In this paper, authors consider *privacy* as one of the most important properties that information systems must mandatorily satisfy. In fact, in these systems there is a need for sharing information among different, untrusted entities, and the protection of *sensitive information* plays a relevant role. A relatively-new trend shows that classical *access control techniques* are not sufficient to guarantee privacy preserving when data mining techniques are used in a malicious way. Hence, *privacy preserving data mining algorithms* have been recently introduced with the aim of preventing the discovery of sensitive information. In the privacy preserving association rule mining algorithm version proposed by authors, beyond to performance and accuracy benefits, flexibility for extension to any number of sites/entities can be achieved without any change in the implementation. In addition to this, any increase in number of sites/entities does not add more time overhead, as all client applications perform the mining process by consuming the same amount of time so that the overhead is in communication time only. Finally, the total bit-communication cost for the proposed algorithm is linearly dependent on the number of sites/entities.

The fifth paper, titled 'Context dependent semantic granularity', by Riccardo Albertoni, Elena Camossi, Monica De Martino, Franca Giannini and Marina Monti, focuses the attention on *efficiently dealing with huge amounts of data* as a fundamental issue for improving accessibility to information resources. In this respect, authors argue that *ontology-driven techniques* are expected to improve the overlap between the *cognitive space* applied by users and the *information space* defined by information providers. Based on this main assertion, authors propose a *powerful method for extracting semantic granularities*, which enable the navigation of a repository according to different levels of abstraction. In the proposed formalisation, granularities are explicitly parameterised on the basis of criteria induced by the context, which greatly improves the method flexibility. Furthermore, the parameterisation assists the user in formulating and refining the browsing criteria. Case studies are described in order to demonstrate how granularities ease the information sources browsing and to illustrate how they may vary according to the context. A validation of the cognitive principles behind the method is presented, together with the analysis of the results obtained by a comprehensive experimentation.

The editor would like to thank very much the editor-in-chief of the *International Journal of Data Mining, Modelling and Management*, Prof. John Wang, for accepting his proposal of a special issue focused on mining, modelling and managing complex data, and for assisting him whenever required. The editor would also like to thank all the reviewers who have worked within a tight schedule and whose detailed and constructive feedbacks to authors have contributed to substantial improvement in the quality of final papers. Last but not least, the editor is grateful to the authors who have submitted papers to this special issue. The editor truly appreciates their patience and understanding throughout the review process.