

Preface

The 25th International Conference on Database and Expert Systems Applications (DEXA 2014), with proceedings published in volumes 8644 and 8645 of Springer's *Lecture Notes in Computer Science*, featured some outstanding keynote presentations and regular articles. As with previous editions of the conference, the program co-chairs of DEXA 2014 invited some of the authors to submit extended papers to a special issue of the Springer journal *Transactions on Large-Scale Data- and Knowledge-Centered Systems* (TLDKS). Following these invitations, one keynote paper and eight regular articles were submitted. Apart from the keynote paper, each submission was carefully assessed by at least two (often more) recognized experts in the respective field. In two rounds of assessment, 35 reviews were received, most of them of very good quality. In the end, six of the eight regular papers were accepted for inclusion in this special issue, in addition to a revised and extended keynote paper.

The contributions in this special issue address a wide range of important contemporary subject areas in data-centric systems and applications, including reflective modeling, big data, similarity search, large-scale data replication, bioinformatic workflows, data pricing, and data anonymization. In good DEXA tradition, all contributions distinguish themselves by the novelty and innovation they bring to these subject areas.

The keynote paper is authored by Dirk Draheim, who, apart from teaching, leads the Information Technology Services Centre at the University of Innsbruck (Austria), and in particular the High-Performance Computing Department of that center. He is a distinguished expert in the field of systems modeling. Theoretical work on modeling usually focuses on logical abstractions of data structures, objects, processes, and object-level language constructs. Additionally, meta-data are of increasing relevance and importance in many fields of information processing. Hence, what is also needed, yet rarely provided, are means to make metadata accessible on the object level. The main contribution of the keynote paper, entitled "Reflective Constraint Writing," with subtitle "A Symbolic Viewpoint of Modeling Languages," consists in a formal treatment of how to add reflection to object-oriented constraint languages. The work presented in this paper is extensive, covering both introspective as well as manipulative data access, for a large variety of purposes, such as to make models robust against unwanted updates, to give precise semantics to existing modeling language constructs, to enable a more adequate system analysis, to assure the quality of system design, and so on.

The article "PPP-Codes for Large-Scale Similarity Searching" is co-authored by David Novak and Pavel Zezula from Masaryk University, Brno, Czech Republic, and supported by the Czech Research Foundations. Their research addresses the challenging problem of efficiently identifying objects in large search spaces that are similar to a given object. The contribution is a two-phase search algorithm on top of a new sophisticated data structure called PPP-Code index. Phase one computes independent rankings based on a given distance function, while phase two aggregates these rankings

to access similar objects sooner. Experiments with artificial and real-world data show that the algorithm reduces the size of output candidates by up to two orders of magnitude while preserving the quality of the answer.

Mouhamadou Ba, Sébastien Ferré, and Mireille Ducassé from IRISA/INSA Rennes and the University of Rennes, France, co-authored the article “Solving Data Mismatches in Bioinformatics by Generating Data Converters.” Their research addresses the prevalent problem where different bioinformatics services with mismatches between given outputs and required inputs need to be composed into workflows. The main contribution is an automatic converter that utilizes a rule-based convertibility detection mechanism. Experiments with real-world data types and services from the bioinformatics domain yielded new composition strategies that domain experts were not made aware of by existing ad-hoc approaches.

The paper entitled “A Framework for Sampling-Based XML Data Pricing” has been written by Ruiming Tang, Antoine Amarilli, Pierre Senellart, and Stephane Bressan. The authors are affiliated to the National University of Singapore, the National Scientific Research Centre in Paris (France), or to both. The paper presents a sharp-witted approach to determining the market value of XML data based on data samples. The price of the data depends on the degree of completeness of sampling and on the contextual quality of data. In other words, data completeness can be traded for a discount price. The paper is one of the first to reflect the growing perception of data as merchandizing objects. In fact, the importance of data pricing is very likely to increase with the expected expansion of electronic information trading. Hence, for future work in that growing trend, this paper can be expected to become a point of reference.

The authors Nikolaos Nodarakis, Evaggelia Pitoura, Spyros Sioutas, Athanasios Tsakalidis, Dimitrios Tsoumakos, and Giannis Tzimas of the paper “*kdANN+: A Rapid kNN Classifier for Big Data*” work at the universities of Patras, Ioannina, the Ionian University in Corfu, and the Technological Educational Institute in Patras, Greece. They propose the use of kNN classification for multidimensional objects. Their paper reports on a novel application in the area of big data, based on the all k-nearest neighbor query method. A divide-and-conquer strategy is pursued: Data space decomposition techniques are deployed for reducing the demand of computational resources. The authors have verified the viability of their solution on experimental data sets. By increasing the dimensionality of the dataset, the total execution cost may exceed the computational power of the cluster infrastructure at hand. To cope with that, the authors propose dimensionality reduction techniques. Their results exhibit differences of computation time and cost between the examined algorithms kdANN and kdANN+, with regard to space dimensionality, the granularity of space decomposition, and the number of nearest neighbors.

The paper entitled “Optimizing Inter-Data-Center Large-Scale Database Parallel Replication with Workload-Driven Partitioning” is authored by Zhen Gao, Hong Min, Xiao Li, Jie Huang, Yi Jin, and An Lei. They are affiliated to various IBM labs in the USA or China, to Tongji University in Shanghai (China), or Pivotal Inc. in Beijing (China). The authors propose two algorithms in order to, firstly, partition a large number of workload-associated tables into a minimal number of point-in-time consistency groups that respect some latency constraints, and, secondly, to refine such partitioning by minimizing the number of transaction splits among the resulting

consistency groups. Point-in-time consistency is an important property of replicated data and a critical objective of distributed database management systems. In particular, it is relevant to the design of data repositories that need to be able to handle unexpected shut-downs, so that replica consistency can be recovered. Given the growing use of replication for handling critical big data management challenges, the potential impact of this paper is obvious.

The paper entitled “Anonymization of Data Sets with NULL Values” is authored by Margareta Ciglic, Johann Eder, and Christian Koncilia, from the Alpen Adria University at Klagenfurt (Austria). The paper deals with the problem of anonymizing data with missing or unknown values. NULL values usually represent epistemic gaps, and, in the context of this paper, should not be confused with values that have been used deliberately in order to anonymize data. A solution to the problem of anonymizing data sets with unknown component values has been missing. Rather, database table rows containing NULL-valued attributes are offhandedly discarded in conventional approaches. That, however, may easily yield a disturbing loss of information, which may distort data analysis results and thus lead to faulty decision making. Thus, the paper meets an evident desirability of solutions that do not ignore NULL values, in particular in the context of large volumes of data with a big informational and structural variety, where NULL values are ubiquitous.

To conclude, we would like to thank all authors for their contributions to this special issue. Also, we are grateful to all reviewers for their invaluable work in assessing the papers, thus contributing to the high quality of this collection of articles. Last, but not least, our gratitude goes to Gabriela Wagner, whose editorial assistance and handling of all the communication with the authors and the reviewers finally made this volume possible.

October 2015

Hendrik Decker
Lenka Lhotska
Sebastian Link

Transactions on Large-Scale Data- and
Knowledge-Centered Systems XXIV
Special Issue on Database- and Expert-Systems
Applications

Hameurlain, A.; Küng, J.; Wagner, R.; Decker, H.;
Lhotska, L.; Link, S. (Eds.)

2016, XI, 221 p. 77 illus. in color., Softcover

ISBN: 978-3-662-49213-0