Editorial to the special issue: "Trends and advances in database systems research"

Theo Härder · Bernhard Mitschang

The BTW conference, organized by the Fachbereich Datenbanken und Informationssysteme (DBIS) of the Gesellschaft für Informatik—GI (German Computer Society), is addressing all subjects in the areas of databases and information systems. It constitutes the central forum of the database communities in Germany, Austria, and Switzerland, where not only scientists but also practitioners and users present their latest results in research and development and exchange information on many topics of database and information technology. The bi-annual BTW started as a conference series for "Datenbanksysteme in Büro, Technik und Wissenschaft" in 1985; at that time, database technology was evolving from its prior focus on storing and maintaining conventional business data towards so-called non-standard database applications which included applications in office automation, engineering, and scientific data management (which explains the German title and also the acronym). Several years ago, it was renamed to reflect the current focus of database management and now stands for "Database Systems for Business, Technology, and Web".

The 14th BTW conference (BTW 2011) took place from February 28th to March 4th, 2011 at the University of Kaiserslautern, Germany. A particular focus of this Technical University is on information technology—backed by several research institutes such as the Fraunhofer Institute for Experimental Software Engineering (IESE), the Institute

for Technical and Business Mathematics (ITWM), and the German Research Center for Artificial Intelligence (DFKI). The recent establishment of the Max-Planck-Institute for Software Systems has further strengthened its position in IT-related research.

Just as the properties of data being managed have changed over the last 25 years, the techniques for data management and processing have also changed and adapted themselves to the new challenges. In addition to the treatment of structured data in past years, semi- or unstructured data gained more and more attention. Classic centralized database system architectures have been extended and partially replaced by main-memory-based, distributed, parallel, or open systems. New hardware technologies are playing an increasingly important role: these include mobile systems, multi-core systems, graphics cards or special storage media such as flash memory. Addressed by the program of the BTW 2011, these techniques are reflected in the current research topics of interest: information integration, data analysis, ontologies and Semantic Web, data processing, service-oriented architecture, cloud computing, virtualization technologies, energy efficiency, and many more.

The BTW 2011 requested papers on core topics (technology) and applications of databases (business and Web) of equal importance. From about 100 submissions, 24 long and 6 short scientific papers were accepted and published in an electronic conference proceeding of the GI-Edition (Lecture Notes in Informatics, Volume P-180) together with three keynote papers or abstracts, the summary of the dissertation selected for the DBIS prize, 9 industry contributions, and 12 descriptions of software demonstrations. The five scientific contributions obtaining the highest reviewer scores were selected for publication in this special issue. For this purpose, the authors were asked to submit improved and extended versions of their papers which were reviewed anew.

T. Härder (✉)
Fachbereich Informatik, TU Kaiserslautern, Kaiserslautern, Deutschland
e-mail: haerder@cs.uni-kl.de

B. Mitschang
Institut für Parallele und Verteilte Systeme, Universität Stuttgart, Stuttgart, Deutschland
e-mail: Bernhard.Mitschang@ipvs.uni-stuttgart.de

The results is this special issue of CSRD which covers a broad spectrum of current research in interesting database topics.

The first paper, *New algorithms for join and grouping operations* by Goetz Graefe (Hewlett-Packard Laboratories, Palo Alto, USA), is related to relational query processing and optimization. For both types of operations, all the numerous proposals concerning their implementation can be classified into nested-loops-based, sort-merge-based, or hash-based algorithms. Industrial-strength solutions in database management systems for joins include the *index nested-loops join* which uses an index for accessing its inner input table, the *merge join* which takes advantage of sorted input tables, and the *hash join* which exploits differences in the table sizes of the join inputs. Grouping operations typically have exploited indexes in the past, but today sort- and hash-based algorithms become widely accepted. As a trend-setting contribution, Goetz Graefe generalizes the algorithms for both types of operations with the goal to keep their strengths and avoid their weaknesses. Hence, the three traditional types of algorithms can be replaced with a single one. As a consequence, cost-based query optimization does not need to choose anymore the most appropriate algorithm for each query and for each operation. Obviously, this approach avoids mistaken algorithm choices during compile-time query optimization for join and for grouping operations.

The paper *Low-overhead decision support for dynamic buffer reallocation*, written by Karsten Schmidt and Sebastian Bächle (University of Kaiserslautern, Germany), deals with self-tuning buffer management in database systems. Optimal use of a buffer configuration—consisting of several buffers managed by different page replacement algorithms—needs dynamic and continuous adjustment of the individual buffer sizes. Keeping the available memory allocation unchanged, reliable forcasts of I/O costs at run-time have to support the decision whether (and how much) the individual buffer sizes should shrink or grow. For this purpose, a prediction model is described whose effectivity is demonstrated by a number of experiments using different page replacement strategies.

The third contribution *Multi-pass sorted neighborhood blocking with MapReduce* by Lars Kolb, Andreas Thor, and Erhard Rahm (University of Leipzig, Germany) explores the use of the MapReduce paradigm for the so-called entity resolution, i.e., to identify different entities in large data sets which represent the same object of the miniworld. Entity resolution is a technique to solve an important problem, e.g., related to the integration of Web data sources. Therefore, suitable infrastructures should enable efficient parallel processing of such data-intensive tasks. Based on the MapReduce programming model, the authors propose two methods for parallel entity resolution using the so-called Sorting Neighborhood blocking. Experiments concerning duplicate detection on real-world datasets confirm the efficiency and effectiveness of the approaches.

The contribution *Conceptual views for entity-centric search: turning data into meaningful concepts* by Joachim Selke, Silviu Homoceanu, and Wolf-Tilo Balke (Technische Universität Braunschweig, Germany) copes with information extraction and tries to close the semantic gap between the information needs of the users and the data represented by unconventional types, often as entity descriptions in unstructured textual form—such as product reviews, user ratings, tags, and images. Conceptual views aim to uncover query-relevant concepts reflected by unstructured data. Two case studies illustrate how meaningful conceptual information can automatically be extracted from existing data, thus enabling the effective handling of vague real-world query concepts.

The final contribution is delivered by Parisa Haghani (EPF Lausanne, Switzerland), Sebastian Michel (Universität des Saarlandes, Germany), and Karl Aberer (EPFL). Their paper *Efficient monitoring of personalized hot news over Web 2.0 streams* describes an approach for efficient evaluation of top-k queries on continuous Web data streams such as blog postings, microblogging tweets, or RSS feeds from on-line communities. Pre-aggregated index lists are used to dynamically reconstruct the results of top-k queries. The suitability of the method is shown by a performance evaluation using a real-world dataset obtained from a weblog crawl.

We want to thank the authors for their timely and smooth cooperation during the preparation of this special issue. Thanks also go to the reviewers who have carefully evaluated and commented the contributions in a short time frame.