The VLDB Journal (2007) 16:1–3
DOI 10.1007/s00778-006-0031-0

GUEST EDITORIAL

Special issue: best papers of VLDB 2005

Laura M. Haas · Christian S. Jensen · Martin L. Kersten

Published online: 13 September 2006

It has become a tradition to dedicate an issue of The VLDB Journal to selected papers from the International Conference on Very Large Databases. This issue is dedicated to the 31st VLDB conference, which was held in Trondheim, Norway from August 30 to September 2, 2005. Following the vision of the VLDB Endowment to further broaden the scope of database research, the conference encompassed four program committees: core database technology, infrastructure for information systems, industrial applications and experiences, and demonstrations.

An estimated 2,000+ authors from virtually all over the world submitted 563 papers. The core database technology committee received 322 submissions, out of which 53 were accepted, the infrastructure for information systems committee received 195 submissions, out of which 32 were accepted, and the industrial application and experiences committee received 46 submissions, out of which 18 were accepted.

Extended and improved versions of seven papers accepted by the core database technology and infrastructure for information systems committees are included in this special issue. These papers were obtained by inviting the authors of the papers considered for the VLDB 2005 Best Paper Award to submit extended versions of their papers. The papers in this special issue are thus the results of the subsequent reviewing and revision of the extended submissions. The issue covers a broad area of database research and highlights the vitality of the research area.

In their paper, Godfrey, Shipley, and Gryz study the problem of efficient maximal-vector computation or, as it is often referred to in a database context, skyline computation. Given a pair of vectors, or tuples, the one vector or tuple dominates the other if for each dimension or attribute, its value is equal to or better than that of the other, and if it has a better value for at least one dimension or attribute. Given a set of vectors, the problem is then to find efficiently all vectors that are not dominated by another vector. This functionality is quite fundamental, and because maximal vectors are intuitively "interesting," it has many uses. The paper offers a wonderful introduction to its topic, it offers new insights into the problem and the behavior of existing solutions, and it proposes a new, efficient solution.

The query optimizer is at the core of a relational database management system, and it is arguably the most complex component of such a system. The optimizer is responsible for turning queries formulated in the high-level language SQL into code that computes the queries efficiently. With the introduction of materialized views, it has become important for optimizers to be able to exploit these well in their efforts to generate efficient code. Larson and Zhou extend techniques for the expression of queries in terms of materialized views, called view matching, to cover views that utilize selection, projection, inner and outer joins, and aggregation while contending with the bag semantics of SQL. Their matching techniques are capable of reasoning about semantic equivalence and subsumption, and they exploit schema

L. M. Haas
IBM Almaden Research Center,
San Jose, CA, USA

C. S. Jensen (✉)
Department of Computer Science,
Aalborg University, Aalborg, Denmark
e-mail: csj@cs.aau.dk

M. L. Kersten
CWI, Amsterdam, The Netherlands

information such as not-null, uniqueness, and foreign-key constraints.

The next paper also concerns an important problem in query optimization. An optimizer generates a number of plans for executing a query and needs to be able to estimate the cost of executing each plan, so that it can choose to actually execute a plan that is efficient. This is done in part by estimating the sizes of intermediate results, as these sizes affect the costs of the intermediate computations. Size estimation in turn translates into selectivity estimation for predicates. In their paper, Markl, Haas, Kutsch, Megiddo, Srivastava, and Tran present selectivity estimation techniques for conjunctive predicates that are based on the maximum entropy principle. Their techniques advance the state of the art by ensuring consistency and by being able to exploit all information available to the query optimizer. They demonstrate that their techniques are capable of better estimates by orders of magnitude and yield better plan selection and thus improved query processing efficiency.

The Best Paper Award for VLDB 2005 was awarded to the paper by Ghoting, Buehrer, Parthasarathy, Kim, Nguyen, Chen, and Dubey. In this pioneering work, the authors examine the performance of mining algorithms on modern hardware, discovering that even the best frequent pattern mining implementations cannot fully exploit a modern processor due to poor data locality and low instruction level parallelism. They propose a new data structure to support mining, called a cache-conscious prefix tree, and a novel tiling strategy, called path tiling, to improve spatial and temporal locality, respectively. Performance studies demonstrate speedups over other mining implementations that exceed a factor of 3. With further enhancements for simultaneously multi-threaded processors, speedups of almost a factor of 5 are reached. Improvements also benefit other architectures.

Schema matching is a critical step in working with data from multiple data sources, needed for information integration, e-commerce, scientific collaboration, as well as other applications. Most of today's schema matching systems utilize several matching techniques, allowing the user to tune the system to select which should be executed and to adjust thresholds, formula coefficients, and so on. The next paper, by Lee, Sayyadian, Doan, and Rosenthal, shows that the accuracy of the matches depends heavily on the tuning, and it introduces an automated tuning method, eTuner. The authors develop methods to tune a broad range of matching components, and use eTuner to tune four matching systems on several real-world domains, demonstrating higher accuracy in matching than possible with current tuning methods.

Data ambiguity arises in many contexts, and the representation and querying of imprecise and uncertain data has been studied extensively. The paper by Burdick, Deshpande, Jayram, Ramakrishnan, and Vaithyanathan concerns multidimensional OLAP over ambiguous data, extending the OLAP data model to represent both imprecision and uncertainty. The authors introduce three criteria – consistency, faithfulness, and correlation preservation – which they use to arrive at an appropriate query semantics based on possible worlds. They present efficient algorithms for evaluating aggregation queries over ambiguous databases, showing that it is possible to evaluate such queries without enumerating all possible worlds, and they also provide complexity analyses. The paper concludes with an empirical evaluation that considers scalability as well as result quality.

The next paper, by Haftmann, Kossmann, and Lo, concerns a very important aspect of software engineering, namely the testing of database applications. The authors note that 50% of Microsoft's development costs are due to testing and that SAP use 6 months out of an 18-month product cycle on testing. As database applications become increasingly complex, they need to be changed more frequently. Testing thus gains further in importance. The authors observe that most of today's tools and techniques for software testing do not contend well with database applications as they are unable to appropriately take into account the database state. To enable efficient regression testing of database applications, the authors propose a testing framework that aims to enable parallel testing and utilize the available system resources. A key aspect is to automatically control the database state so that expensive state reset operations are needed only rarely. Performance studies indicate that the framework is indeed able to yield linear speedups in the number of parallel machines used.

The technical program of VLDB 2005, from which the papers in this special issue are drawn, resulted from the combined efforts of a large team. Per-Åke Larson chaired the industrial applications and experiences program committee, and Beng Chin Ooi chaired the demonstrations program committee. The four program committees had some 171 members, each of whom typically was responsible for reviewing a dozen papers. In doing so, they solicited the help of some 440 external referees. Klaus Dittrich coordinated the affiliated workshops, Tore Risch handled panels, Masaru Kitsuregawa was in charge of tutorials, Betty Salzberg chaired the program committee of the affiliated Ph.D. workshop, and Øystein Torbjørnsen handled exhibitions. We thank them all for putting their expertise to work on the task of creating an exciting technical program. Klemens Böhm was in charge of compiling the proceedings – we thank

him and his team for doing an expert job. Finally, we extend our thanks to the team of local organizers headed by Kjell Bratbergsengen and Mads Nygård. It has been a pleasure working with them.

Last but not least, we thank those who reviewed the extended versions that were considered for inclusion in this special issue.