Li GL, Tang N, Chai CL. Special section on scalable data science. JOURNAL OF COMPUTER SCIENCE AND TECHNOLOGY 37(5): 1003–1004 Sept. 2022. DOI 10.1007/s11390-022-0005-8

Preface

Data science targets the data life cycle of real applications, studying phenomena at scales, complexities, and granularities never before possible. This data life cycle encompasses databases and data engineering often leveraging statistical, machine learning, and artificial intelligence methods and, in many instances, using massive and heterogeneous collections of potentially noisy datasets.

To promote the recent work on scalable data science, we organize this special section at Journal of Computer Science and Technology (JCST). In this special section, we focus on data-intensive components of data science pipelines, and solve problems in areas of interest to our community (e.g., data curation, optimization, performance, storage, and systems). We received 29 papers from all over the world. First, the guest editors preformed quick reviews and immediately rejected insufficiently high quality submissions. Then, each remaining submission was reviewed by at least three invited international reviewers. All the papers were carried out two rounds of reviews, and the authors were asked to address all the major and minor issues in their submissions during the review process. Eventually we accepted six high-quality submissions in terms of clarity, novelty, significance, and relevance for this special section.

The first paper "GAM: A GPU-Accelerated Algorithm for MaxRS Queries in Road Networks" by Chen *et al.* proposes a novel GPU-accelerated algorithm GAM to tackle maximizing range sum queries in road networks efficiently with a two-level framework. The framework first proposes an effective multi-grained pruning technique to prune the cells derived from partitioning the road network, and then a GPU-friendly storage structure is designed to compute the final result in the remaining cells.

The second paper "Experiments and Analyses of Anonymization Mechanisms for Trajectory Data Publishing" by Sun *et al.* systematically evaluates the individual privacy in terms of unicity and the utility in terms of practical applications of the anonymized trajectory data. This paper reveals the true situation of the privacy preservation for trajectories in terms of reidentification and the true situation of the utility of anonymized trajectories.

The third paper "Efficient Partitioning Method for Optimizing the Compression on Array Data" by Han *et al.* utilizes header compression to address the problem of array partitioning for optimizing the compression performance. The paper designs a greedy strategy which can help to find the partition point with the best compression performance.

The forth paper "Discovering Cohesive Temporal Subgraphs with Temporal Density Aware Exploration" by Zhu *et al.* proposes a temporal subgraph model to discover cohesive temporal subgraphs by capturing both the structural and the temporal characteristics of temporal cohesive subgraphs. This paper designs strategies to mine temporal densest subgraphs efficiently by decomposing the temporal graph into the sequence of snapshots.

The fifth paper "Incremental User Identification Across Social Networks Based on User-Guider Similarity Index" by Kou *et al.* proposes an incremental user identification method across social networks based on User-guider Similarity Index. The paper first constructs a novel user-guider similarity index to speed up the matching between users, and then applies a two-phase user identification strategy to efficiently identify users.

The sixth paper "An Exercise Collection Auto-Assembling Framework with Knowledge Tracing and Reinforcement Learning" by Zhao *et al.* introduces an exercise collection auto-assembling framework, in which the assembled exercise collection can meet the teacher's requirements on the difficulty index and the discrimination index. The paper designs a two-stage approach where a knowledge tracing model is used to predict the students' answers and a deep reinforcement learning model to select exercises to satisfy the query parameters.

We thank all the authors who submitted to this special section, and are grateful to the reviewers who provided valuable review feedback. We hope that readers will enjoy this special section.

Leading Editor

Guo-Liang Li (李国良), Professor, Department of Computer Science and Technology, Tsinghua University, Beijing
liguoliang@tsinghua.edu.cn

Guest Editors

Nan Tang (汤　南), Senior Scientist, Qatar Computing Research Institute, Hamad Bin Khalifa University, Doha
ntang@hbku.edu.qa

Cheng-Liang Chai (柴成亮), Postdoctoral Researcher, Department of Computer Science and Technology, Tsinghua University, Beijing　ccl@tsinghua.edu.cn

Guo-Liang Li is a full professor at Department of Computer Science and Technology, Tsinghua University, Beijing. He got his Bachelor's degree in computer science from Harbin Institute of Technology, Harbin, in 2004, and his Ph.D. degree in computer science from Tsinghua University, Beijing, in 2009. His research interests include AI4DB, DB4AI, big data management, crowdsourced data management, and large-scale data cleaning and integration. He got VLDB 2017 Early Research Contribution Award, TCDE 2014 Early Career Award, CIKM 2017 Best Paper Award, VLDB 2020 Best Papers, KDD 2018 Best Papers, ICDE 2018 Best Papers, DASFAA 2014 Best Paper Runnerup, APWeb 2014 Best Paper Award, and EDBT 2013 Similarity Join and Search Champion. He was the general chair of SIGMOD 2021 and regularly served as a PC member of SIGMOD, VLDB, ICDE, KDD, and WWW. He is serving as an associate editor for IEEE TKDE, VLDB Journal, etc.

Nan Tang received his Ph.D. degree from The Chinese University of Hong Kong, Hong Kong, in 2007. He is a senior scientist at Qatar Computing Research Institute, Hamad Bin Khalifa University, Doha. He has worked as a research staff member at Centrum Wiskunde & Informatica (CWI), The Netherlands, from 2008 to 2010. He was a research fellow at the University of Edinburgh, Edinburgh, from 2010 to 2012. His current research interests include data curation and data streams.

Cheng-Liang Chai is a postdoctoral researcher at Department of Computer Science and Technology, Tsinghua University, Beijing. He received his Ph.D. degree in computer science from Tsinghua University, Beijing, in 2020. He received the 2020 ACM China Doctoral Dissertation Award and the 2020 CCF Doctoral Dissertation Award. His research interests include leveraging data management techniques to benefit artificial intelligence, including data cleaning, data discovery and labeling and utilizing artificial intelligence to improve the database performance, including the learned optimizer, learned index and database tuner.