Distributed and Parallel Databases (2022) 40:201–204
https://doi.org/10.1007/s10619-022-07420-y

Introduction to special issue on scientific and statistical data management in the age of AI 2021

Qiang Zhu[1] · Xingquan Zhu[2] · Yicheng Tu[3]

Published online: 22 August 2022

Recent advancement in deep neural networks, combined with the high performance computing power and Big Data, has profoundly brought artificial intelligence (AI) to nearly all fields of scientific disciplines. Data analytics and interpretation play an essential role in scientific studies. From molecular dynamics, next-generation sequencing, to the study of environments and universe such as ocean current dynamics and supernova explosion, data science is often at the frontier of these studies. During the Big Data era starting in early 2000, scientific and statistical data management primarily focused on data harvesting, such as efficient transportation, compression, storage, and querying of large amounts of data. For example, distributed file systems such as HDFS and in-memory MapReduce like Spark have been introduced to make large scale data management and processing possible. This development essentially laid the foundation for our currently flourishing cloud computing systems such as AWS. With the advancement in deep learning and other machine learning (ML) technologies, combined with GPU and TPU powered computing capacities, the attention of data management has gradually shifted to focus more on promoting understanding and use of patterns, trends, and knowledge underneath data, especially dealing with large amounts of data, beyond traditional data management areas. As ones have witnessed, machine learning based systems are becoming rapidly available in ecommerce, autonomous driving, robotics, navigations, and many other areas. There is an increasing demand to better integrate data management with AI/ML to provide seamless solutions for the complete process of data preparation,

✉ Qiang Zhu
qzhu@umich.edu

Xingquan Zhu
xzhu3@fau.edu

Yicheng Tu
tuy@mail.usf.edu

[1] University of Michigan – Dearborn, Dearborn, MI, USA

[2] Florida Atlantic University, Boca Raton, FL, USA

[3] University of South Florida, Tampa, FL, USA

integration, management, analytics, and interpretation to benefit scientific studies and others.

The 33rd International Conference on Scientific and Statistical Database Management (SSDBM 2021), July 6–7, 2021, brought together scientific domain experts, database researchers, practitioners, and developers for the presentation and exchange of current research results on concepts, tools, and techniques for scientific and statistical database applications. 59 papers, submitted by 243 authors from over 20 countries/regions to SSDBM 2021, have showcased the pervasiveness of AI/ML in the field of scientific and statistical data management, which covered a variety of related topics (among others) including:

- In-database machine learning.
- Data mining for graphic/time series/spatial/temporal/senor/streaming/trajectory/mobile/real-time/social network/privacy-preserving/biomedical/scientific data.
- Deep learning for anomaly detection.
- Meta-data modeling for statistical data.
- k-NN and similarity searches.
- NLP over knowledge graphs of scientific data.
- Ontology and semantic web.
- Recommender systems.
- Resource prediction.
- Intelligent query processing.
- Data quality and visualization.
- Data preparation/integration for AI applications in array/main-memory/GPU-accelerated/personal/cloud/distributed databases.

After careful reviews and discussions, the SSDBM 2021 Program Committee accepted 16 full research papers, 13 short research papers, and one demonstration paper for the program of the conference.

As one of the postconferene initiatives, we are pleased to present this special issue of the Parallel and Distributed Databases (DAPD) Journal on *Scientific and Statistical Data Management in the Age of AI 2021* that features the extended versions of six papers selected from SSDBM 2021 after a rigorous peer review process. In the following, we provide a brief overview of the six papers in this special issue.

The paper entitled "Recursive SQL and GPU-Support for In-Database Machine Learning" (by Maximilian E. Schule, Harald Lang, Maximilian Springer, Alfons Kemper, Thomas Neumann, and Stephan Gunnemann) presents a novel study showing that SQL with recursive tables, data sampling, and continuous views can formulate a complete machine learning pipeline including data preprocessing, model training, and model validating. A database operator for automatic differentiation using a lambda expression as well as a dedicated operator for gradient descent using recursive computation are introduced, which allow expressing broad loss functions including matrix operations as used for a neural network. The training algorithms are implemented as GPU kernels to offload training to GPU units. These kernels are integrated inside the code-generating database system Umbra. This work has made a significant contribution to the area of in-database learning, especially, introducing

new machine learning operators in SQL and improving overall performance utilizing available hardware.

The paper entitled "ReSKY: Efficient Subarray Skyline Computation in Array Databases" (by Dalsu Choi, Hyunsik Yoon, and Yon Dohn Chun) studies a new type of subarray skyline query for array databases, which have become popular in managing large-scale spatial data generated in fields including scientific studies and location-based services. These subarray skyline queries are useful for applications where a change typically affects a set of neighboring cells (subarray) together for an underlying array database. A subarray skyline is defined by the aggregate characteristic values for subarrays. To efficiently process subarray skyline queries, the authors adopt intelligent search strategies to prune non-skyline subarrays early and reduce unnecessary dominance checks among skyline subarrays. To support large-scale spatial data, the authors apply global sampling and load balancing strategies to extend their centralized subarray skyline query processing method to a distributed coordinator-worker environment.

The paper entitled "Scalable Probabilistic Truss Decomposition using Central Limit Theorem and H-Index" (by Fatemeh Esfahani, Mahsa Daneshmand, Venkatesh Srinivasan, Alex Thomo, and Kui Wu) investigates intelligent computing methods and their theoretical foundation for truss decomposition in probabilistic graphs. Truss decomposition is a popular notion to extract hierarchical dense substructures in graphs. It has various useful applications such as network robustness analysis, complex network visualization, social community modeling, and genome-based disease discovery. To tackle the challenges for computing truss decomposition in large probabilistic graphs, the paper suggests a novel approach based on Lyapunov's Central Limit Theorem to approximate the probability distribution of the support of an edge. Using this fast calculation of edge support probabilities and optimized array-based data structures, the paper proposes an efficient peeling algorithm for computing truss decomposition in large probabilistic graphs. To achieve small memory footprint and progressive computation, the paper also introduces another h-index based algorithm for computing truss decomposition for large probabilistic graphs. In addition, interesting discussions on the approximation error bound, iteration upper bound, algorithm correctness, and computational complexity are given.

The paper entitled "Four Node Graphlet and Triad Enumeration on Distributed Platforms" (by Yudi Santoso, Xiaozhou Liu, Venkatesh Srinivasan, and Alex Thomo) presents a study on distributed graphlet enumeration for large input graphs. Many applications in biology, chemistry, social study, network analysis and classification, and so on need graphlet enumeration for large graphs. Existing solutions suffer major limitations on the scale of an input graph and the order of a graphlet that they can efficiently handle. To mitigate the problem, this paper presents an efficient distributed algorithm for enumerating all induced four-node graphlets in undirected graphs on a single run. Intelligent strategies to suppress duplicate computation are adopted. Detailed analyses of the correctness and efficiency along with empirical evaluation are discussed. The paper also introduces an efficient distributed triad enumeration algorithm with demonstrated good scalability for directed graphs.

The paper entitled "Structured Data Transformation Algebra (SDTA) and Its Applications" (by Jie Song, George Alter, and H. V. Jagadish) introduces a formal

paradigm for statistical data transformation, which includes an algebra with operators covering essentials of statistical transformation as well as a generic data model allowing information propagation among a transformation flow. Statistical data transformation is a critical component of many data analytic processes in data science. It is generally accomplished by using scripts in various statistical languages (e.g., SPSS, Stata, and SAS) based on disparate data models. The paper presents a generic declarative statistical data transformation language based on the proposed algebra. It shows that statistical data transformation scripts in various proprietary statistical languages can be converted into equivalent scripts in the proposed algebra-based generic language. With this approach, not only metadata including provenance information can be properly documented, but also code reuse, result reproducibility, and rewriting-rule-based transformation optimization can be offered.

The paper entitled "Bio-SODA UX: Enabling Natural Language Question Answering over Knowledge Graphs with User Disambiguation" (by Ana Claudia Sima, Tarcisio Mendes de Farias, Maria Anisimova, Christophe Dessimoz, Marc Robinson-Rechavi, Erich Zbinden, and Kurt Stockinger) aims at tackling the challenges of natural language processing over knowledge graphs of scientific datasets where no prior training data is available. The paper introduces a natural language processing system designed to answer natural language questions across knowledge graphs for scientific data applications where no prior training data in the form of question–answer pairs is available. It employs a generic graph-based approach to translate natural language questions into SPARQL candidate queries. An effective ranking method that takes into consideration syntactic and semantic similarity as well as node centrality is adopted to select the best SPARQL candidate query. Furthermore, a graphic user interface is designed to help users explore large knowledge graphs and dynamically disambiguate natural language questions.

We wish to thank all the authors for their valuable contributions to this issue. We would also like to thank the reviewers for their thoughtful comments and suggestions that have helped the authors significantly improve their submissions during the review process. Finally, we are grateful to the DAPD Journal editors-in-chief, editorial office, and production team for their tremendous support and assistance in planning, preparing and producing this issue. We hope readers find this special issue interesting and inspiring.