Contents lists available at ScienceDirect

Data & Knowledge Engineering

journal homepage: www.elsevier.com/locate/datak

Preface

Data engineering and modeling for artificial intelligence

1. Introduction

Artificial Intelligence (AI) is now the main goal of practically any analytic project. The growing interest in AI stems from its potential to revolutionize various aspects of business and technology. AI systems can automate routine tasks, provide deep insights through data analysis, enhance decision-making processes, and create personalized user experiences. On the other hand, there is a new trend emerging that focuses on building systems and novel approaches to enable analysis on practically any kind of data. This trend, often referred to as the new generation of Big Data, encompasses theory and technology that can handle data of any size, content and format. Whether a data set is big or small, unstructured or structured, static or streaming, these new systems are designed to process and analyze it effectively. For instance, past database systems might struggled with unstructured data like text, images, or videos, but modern Big Data systems, frameworks and libraries can store, manage and analyze such data types with little effort. Similarly, real-time data streaming from Internet of Thing devices requires specialized tools for timely processing and analysis, which are now becoming more accessible and sophisticated.

Data science is the interdisciplinary approach that facilitates this comprehensive data analysis. It combines expertise from statistics, computer science, domain knowledge, and data engineering to extract meaningful insights from data. The data science workflow typically begins with detecting data quality problems. This is crucial because accurate analysis depends on the reliability of the data. Data scientists must identify and rectify issues such as missing values, outliers, and inconsistencies to ensure the dataset is robust. Due to these limitations, a significant amount of time is spent on data pre-processing. This stage involves cleaning the data, transforming it into a suitable format, and integrating it from various sources. Data pre-processing is essential for preparing the raw data for analysis, and it often involves tasks like normalization, feature extraction, and dimensionality reduction. Once the data is prepared, data scientists apply sophisticated machine learning models to analyze it. Machine learning models can vary widely in complexity, from simple linear regressions to advanced neural networks. Neural networks, in particular, are a common choice for tasks requiring high levels of abstraction and pattern recognition, such as image and speech recognition. These models can learn from the data, identifying trends and making predictions that can inform business strategies and operational decisions.

At the beginning of the explosion of AI-driven solutions, the database community focused on revisiting traditional database problems (e.g., physical design, cleaning, and data placement) and proposing AI-driven techniques to solve them. Following that, the community began exploring how traditional database solutions could enhance AI techniques to satisfy non functional requirements (e.g., scalability, usability, manageability, developability).

With this motivation in mind, we launched the first IEEE International Workshop on Data Engineering and Modeling for Artificial Intelligence (DEMAI 2023), sponsored by the IEEE International Conference on Big Data. Our workshop was held in Sorrento, Italy, on December 15, 2023. This workshop aimed to bridge the gap between research on data management and artificial intelligence. The goal of the workshop was to bring together researchers, analysts, and data scientists to identify research issues and share experiences to integrate and pre-process data before AI models can be computed.

The intersection of AI and advanced data analysis techniques is transforming the landscape of data analytics. By enabling the analysis of diverse and complex data types, these technologies are expanding the possibilities for innovation and efficiency in various fields. The interdisciplinary nature of data science, combining rigorous analytical methods with practical applications, ensures that organizations can harness the full potential of their data. As these trends continue to evolve, they promise to drive significant advancements in technology and industry, ultimately enhancing our ability to understand and interact with the world around us.

https://doi.org/10.1016/j.datak.2024.102346

2. Selected papers

This special issue contains selected papers from DEMAI 2023. The DEMAI workshop attracted 13 submissions from which 6 were accepted as short 5-page papers in IEEE format. Given the selective acceptance rate and papers high quality, all 6 accepted papers were invited for journal submission. All papers went through a rigorous submission and review process. We asked authors to go deeper in theory aspects, rethink their title and expand each paper with at least 50% new material (beyond the 30% threshold imposed by the DKE journal), including theory and experimental evaluation. Moreover, authors were asked to position their paper after the intersection of databases, AI and data science. In the end, after a strict editor screening, and a new round of reviews, including old and new reviewers, we accepted 3 papers, which we summarize below.

- The first article [1], titled "An Interactive Approach to Semantic Enrichment with Geospatial Data", explores the capabilities of SemTUI, a comprehensive framework designed to support the enrichment of tabular data by leveraging semantics and user interaction. Utilizing SemTUI, an iterative and interactive approach is proposed to enhance the flexibility, usability and efficiency of geospatial data enrichment. Using a real-world scenario involving the analysis of kindergarten accessibility within walking distance, the study demonstrates the proficiency of SemTUI in generating precise and semantically enriched location data.
- The second article [2], titled "To Prompt or Not To Prompt: Navigating the Use of Large Language Models for Integrating and Modeling Heterogeneous Data" demonstrates the capability of Large Language Models (LLM's) to effectively extract data from unstructured sources and highlights their potential to streamline data extraction and resolution processes. The paper underscores the necessity of preliminary data modeling decisions to ensure the success of such technological applications. By merging human expertise with LLM-driven automation, this study advocates for the further exploration of semi-autonomous data engineering pipelines.
- Finally, the third article [3], titled "Hermes, a low-latency transactional storage for binary data streams from remote devices" presents a solution that enables both high ingestion rates with transactional data persistence and near real-time, low-latency access to the stream during collection. The proposed solution is particularly suitable for binary data sources such as audio and video recordings in surveillance systems, and it can be extended to various big data scenarios via well-defined general interfaces. Preliminary results obtained with Apache Kafka and MongoDB replica sets show that the proposed solution provides up to 3 times higher throughput and 2.2 times lower latency compared to standard multi-document transactions.

We believe these three papers digest the landscape of how database systems help AI. We hope DKE readers will find the content of this special issue interesting and that it will inspire them to look further into the challenges that are still ahead in applying database knowledge in AI and how AI can optimize databases.

Acknowledgments

We would like to thank all the authors who extended their workshop papers for this special issue. In addition, we are grateful for the support of all reviewers who ensured high quality. We want to thank Tony Hu for hosting our workshop at IEEE Big Data. We thank Carson Wu who gave us direction on where the field of data modeling is going.

References

[1] Roberto Avogadro, Emil Hristov, Milena Borukova, Dessislava Petrova-Antonova, Flavio De Paoli, Michele Ciavotta, Iva Krasteva, An interactive approach to semantic enrichment with geospatial data, Data Knowl. Eng. (2024).

[2] Yasmina Hobeika, Adel Remadi, Karim El Hage, Francesca Bugiotti, To prompt or not to prompt: Navigating the use of large language models for integrating and modeling heterogeneous data, Data Knowl. Eng. (2024).

[3] Daniele Apiletti, Gabriele S. Militone, Giovanni Malnati, Hermes, a low-latency transactional storage for binary data streams from remote devices, Data Knowl. Eng. (2024).

Carlos Ordonez is a professor at the University of Houston. His research is centered on large-scale data science, parallel database systems and big data infrastructure. Carlos got a Ph.D. degree in Computer Science from the Georgia Institute of Technology, USA, in 2000. From 2001 to 2006 Carlos worked on extending the Teradata parallel DBMS with machine learning models and cube techniques. Then in 2006 Carlos joined the University of Houston, where he conducts research on parallel and large-scale data processing systems. Carlos was a visiting researcher at MIT during Summer from 2014 to 2016, working on array and columnar database systems. He worked research scientist at ATT Labs from 2014 to 2015, focusing on the R language and data warehousing.

Wojciech Macyna received his MS degree in Computer Science from Wroclaw University of Technology in 1997 and his Ph.D. degree in 2004. He has worked since 2009 as an assistant professor in the Department of Fundamentals of Computer Science at Wroclaw University of Technology in Poland. Prior to this, he worked for leading Polish IT companies. His current research interests mainly include database systems, data science, and new hardware technologies.

Ladjel Bellatreche is a Distinguished Professor at the National Engineering School for Mechanics and Aerotechnics (ISAE-ENSMA), Poitiers – France since September 2010. From 2012 to 2022, he led the Data and Model Engineering Team at the Laboratory of Computer Science and Automatic Control for Systems (LIAS). Before joining ISAE-ENSMA, Ladjel spent eight years as an Assistant and later an Associate Professor at Poitiers University, France. Throughout his academic career, Ladjel has held various international positions, including visiting Professor roles at the University of New South Wales in Sydney, Australia, in 2015, and at the Université du Québec en Outaouais, Canada, in 2009. He has also been a Visiting Researcher at the Department of Computer Science, Purdue University, USA, in 2001, and at the Department of Computer Science of Hong Kong University of Science and Technology (HKUST), China, from 1996 to 1999. Ladjel's primary research interests encompass various facets of the field, including Data Science, Knowledge Graphs, Big Data Warehouses, Scalable and Parallel Algorithms for Big Data Analytics, Green Analytics, and Software Engineering.

Carlos Ordonez
University of Houston, USA

Wojciech Macyna *
Wroclaw University of Technology, Poland

Ladjel Bellatreche
ISAE-ENSMA, France

* Corresponding editor.