1st International Workshop on Tabular Data Analysis (TaDA)

Vasilis Efthymiou[1], Sainyam Galhotra[2], Oktie Hassanzadeh[3], Ernesto Jiménez-Ruiz[4] and Kavitha Srinivas[3]

[1]Foundation for Research and Technology - Hellas, Greece
[2]Cornell University, USA
[3]IBM Research, USA
[4]City, University of London, United Kingdom

Abstract

With the advent of data lakes and open data repositories containing heterogeneous collections of structured datasets, there is an increasing need for automated methods to analyze tabular data collections for a wide range of applications in data management, data science, and decision support. Our goal in this workshop was to bring together researchers and practitioners working on building such tabular data analysis solutions. TaDa workshop aimed to provide a venue for the growing number of researchers in data management, AI, and Semantic Web communities working on a wide range of problems relevant to tabular data analysis. The first edition of the workshop included two keynote talks, a research track comprising presentations and posters, and invited posters and virtual talks of the work done in these communities.

1. Introduction

Data Analysis, as a crucial process in various domains, involves examining, cleaning, transforming, and modeling data to extract valuable insights, make informed conclusions, and facilitate decision-making [1]. However, performing such data analysis tasks becomes exceedingly complex when dealing with vast and diverse collections of tabular data, commonly found in enterprise data lakes and on the Web. Consequently, this challenge has piqued the interest of researchers and practitioners in data management, AI, and related communities [2, 3, 4, 5, 6].

To address the fundamental research challenges posed by tabular data analysis and foster the development of automated solutions, Tabular Data Analysis (TaDA 2023) workshop (https://tabular-data-analysis.github.io/tada2023/) was organized with the primary goal of bringing together experts from diverse communities. This workshop aimed to create a collaborative environment for researchers and practitioners in data management and AI fields, enabling them to share insights, methodologies, and advancements in tackling the complexities of analyzing large and heterogeneous collections of tabular data. The workshop provided a forum for:

- Exchange of ideas between two communities: 1) an active community of data management researchers working on data integration, schema and data matching problems over tabular data, and 2) a vibrant community of researchers in AI and Semantic Web working on matching tabular data to Knowledge Graphs as a part of the ISWC SemTab Challenge [7, 8, 9, 10].
- Presentation of late-breaking results related to several emerging research areas such as table representation learning and its applications, automation of data science pipelines, and data lake and data lakehouse solutions.
- Discussion of real-world data management challenges related to implementing industrial scale tabular data analysis solutions.

2. Overview of the Program

The workshop received several interesting submissions on the different aspects of tabular data analysis, and each submission was reviewed by at least three reviewers. The accepted papers encompassed a wide range of topics, including data discovery, semantic table understanding, column annotation, and knowledge graph-based techniques. The workshop program consisted of keynote talks from two well-known researchers from the field: Renée Miller from Northeastern University and Alon Halevy from Meta AI.

Renée discussed the advances in tools for data scientists to discover useful tables and pointed out that sophisticated methods to integrate discovered tables are underexplored. She presented two of her recent papers: ALITE [11], a method for integrating tables using full disjunction, and DIALITE [12], an open discovery system for analyzing tables, sharing new benchmarks for evaluation. She also presented open problems and challenges

Joint Workshops at 49th International Conference on Very Large Data Bases (VLDBW'23) — TaDA'23: Tabular Data Analysis Workshop, August 28 - September 1, 2023, Vancouver, Canada

vefthym@ics.forth.gr (V. Efthymiou); sg@cs.cornell.edu (S. Galhotra); hassanzadeh@us.ibm.com (O. Hassanzadeh); ernesto.jimenez-ruiz@city.ac.uk (E. Jiménez-Ruiz); kavitha.srinivas@ibm.com (K. Srinivas)

CEUR Workshop Proceedings (CEUR-WS.org)

in developing and evaluating scalable table search and integration methods on real data.

Alon's keynote emphasized the significance of understanding how individuals can leverage their generated data to enhance their health, vitality, productivity, and overall well-being. He motivated the research on fusing personal digital data, discussed potential pitfalls, and explored multiple approaches to querying timelines. This application area necessitated careful consideration of language models to effectively query partially structured and unstructured data.

Acknowledgements

We would like to thank the steering committee, the program committee, the keynote speakers, and the authors for their contributions. Finally, we thank the workshop attendees for making TaDA a great venue to discuss the works in the area of tabular data analysis.

References

[1] M. Brown, Transforming Unstructured Data into Useful Information, 2014, pp. 211–230. doi:10.1201/b16666-11.

[2] O. Hassanzadeh, A. Kementsietsidis, B. Kimelfeld, R. Krishnamurthy, F. Ozcan, I. Pandis, Next generation data analytics at IBM research, Proc. VLDB Endow. 6 (2013) 1174–1175. doi:10.14778/2536222.2536246.

[3] S. Galhotra, A. Fariha, R. Lourenço, J. Freire, A. Meliou, D. Srivastava, Dataprism: Exposing disconnect between data and systems, in: Z. Ives, A. Bonifati, A. E. Abbadi (Eds.), SIGMOD '22: International Conference on Management of Data, Philadelphia, PA, USA, June 12 - 17, 2022, ACM, 2022, pp. 217–231. doi:10.1145/3514221.3517864.

[4] M. Helali, E. Mansour, I. Abdelaziz, J. Dolby, K. Srinivas, A scalable automl approach based on graph neural networks, Proc. VLDB Endow. 15 (2022) 2428–2436. doi:10.14778/3551793.3551804.

[5] F. Özcan, C. Lei, A. Quamar, V. Efthymiou, Semantic enrichment of data for AI applications, in: M. Boehm, J. Stoyanovich, S. Whang (Eds.), Proceedings of the Fifth Workshop on Data Management for End-To-End Machine Learning, In conjunction with the 2021 ACM SIGMOD/PODS Conference, DEEM@SIGMOD 2021, Virtual Event, China, 20 June, 2021, ACM, 2021, pp. 4:1–4:7. doi:10.1145/3462462.3468881.

[6] F. Nargesian, K. Q. Pu, E. Zhu, B. G. Bashardoost, R. J. Miller, Organizing data lakes for navigation, in: D. Maier, R. Pottinger, A. Doan, W. Tan, A. Alawini, H. Q. Ngo (Eds.), Proceedings of the 2020 International Conference on Management of Data, SIGMOD Conference 2020, online conference [Portland, OR, USA], June 14-19, 2020, ACM, 2020, pp. 1939–1950. doi:10.1145/3318464.3380605.

[7] E. Jiménez-Ruiz, O. Hassanzadeh, K. Srinivas, V. Efthymiou, J. Chen (Eds.), Proceedings of the Semantic Web Challenge on Tabular Data to Knowledge Graph Matching co-located with the 18th International Semantic Web Conference, SemTab@ISWC 2019, Auckland, New Zealand, October 30, 2019, volume 2553 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2020.

[8] E. Jiménez-Ruiz, O. Hassanzadeh, V. Efthymiou, J. Chen, K. Srinivas, V. Cutrona (Eds.), Proceedings of the Semantic Web Challenge on Tabular Data to Knowledge Graph Matching (SemTab 2020) co-located with the 19th International Semantic Web Conference (ISWC 2020), Virtual conference, November 5, 2020, volume 2775 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2020.

[9] E. Jiménez-Ruiz, V. Efthymiou, J. Chen, V. Cutrona, O. Hassanzadeh, J. Sequeda, K. Srinivas, N. Abdelmageed, M. Hulsebos, D. Oliveira, C. Pesquita (Eds.), Proceedings of the Semantic Web Challenge on Tabular Data to Knowledge Graph Matching co-located with the 20th International Semantic Web Conference (ISWC 2021), Virtual conference, October 27, 2021, volume 3103 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2022.

[10] V. Efthymiou, E. Jiménez-Ruiz, J. Chen, V. Cutrona, O. Hassanzadeh, J. Sequeda, K. Srinivas, N. Abdelmageed, M. Hulsebos (Eds.), Proceedings of the Semantic Web Challenge on Tabular Data to Knowledge Graph Matching, SemTab 2022, co-located with the 21st International Semantic Web Conference, ISWC 2022, Virtual conference, October 23-27, 2022, volume 3320 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2023.

[11] A. Khatiwada, R. Shraga, W. Gatterbauer, R. J. Miller, Integrating data lake tables, Proc. VLDB Endow. 16 (2022) 932–945.

[12] A. Khatiwada, R. Shraga, R. J. Miller, DIALITE: discover, align and integrate open data tables, in: S. Das, I. Pandis, K. S. Candan, S. Amer-Yahia (Eds.), Companion of the 2023 International Conference on Management of Data, SIGMOD/PODS 2023, Seattle, WA, USA, June 18-23, 2023, ACM, 2023, pp. 187–190. doi:10.1145/3555041.3589732.