Big Data Integration for Data-Centric AI

Sonia Bergamaschi[1,2], Domenico Beneventano[1], Giovanni Simonini[1], Luca Gagliardelli[1], Adeel Aslam[1], Giulio De Sabbata[1] and Luca Zecchini[1]

[1] *Università degli Studi di Modena e Reggio Emilia, Modena, Italy, <name.surname>@unimore.it*

[2] *CINI Big Data National Lab*

Abstract

Big data integration represents one of the main challenges for the use of techniques and tools based on Artificial Intelligence (AI) in several crucial areas: eHealth, energy management, enterprise data, etc. In this context, Data-Centric AI plays a primary role in guaranteeing the quality of the data on which these tools and techniques operate. Thus, the activities of the Database Research Group (DBGroup) of the "Enzo Ferrari" Engineering Department of the University of Modena and Reggio Emilia are moving in this direction. Therefore, we present the main research projects of the DBGroup, which are part of collaborations in various application sectors.

Keywords

Big Data Integration, Entity Resolution, Data-Centric AI

DBGroup Research Activities

Big data integration is the main research area of the Database Research Group (DBGroup[1]) of the "Enzo Ferrari" Engineering Department of the University of Modena and Reggio Emilia, led by Prof. Sonia Bergamaschi. DBGroup has been working on it for over twenty years, having focused for a long time on the MOMIS (Mediator Environment for Multiple Information Sources) data integration system [1, 2]. An open-source version of MOMIS is currently maintained by DataRiver[2], originally founded as a spin-off of the DBGroup in 2009.

With the ever-increasing importance of Artificial Intelligence (AI), the DBGroup research has been focusing on algorithms and techniques for improving the quality and integration of the data, hence of the AI using it—i.e., *Data-Centric AI*[3]. In the field of big data integration, we have been able to develop several innovative projects, with leading international research collaborations [3, 4, 5, 6, 7], while always paying attention to the possible concrete applications of these innovations [8, 9].

Challenges for Entity Resolution

In recent years, the DBGroup has been active in devising Entity Resolution (ER) algorithms and tools. ER is a fundamental task for big data integration: it aims at identifying different mentions of the same real-world entities in data collections—hence, it is the catalyst for combining information coming from different sources. ER is a complex task and has many pain points: *(i)* it is hard to scale; *(ii)* it is hard to automate with machine learning (ML), since it is hard to gather labelled data; *(iii)* it is incredibly challenging if privacy is a concern—e.g., imagine to look for records about the same person without knowing her name because ananymized. At DBGroup, we have been working on solutions for all these aspects of ER [5, 10]. We devised state-of-the-art methods where ML is employed as a tool to automate ER and to achieve high-quality results [7]. Yet, even if the process is automated, sometimes performing ER on the entire dataset can be unfeasible in the big data context. Hence, we also devised algorithms and a tool [6] for combining the ER within SQL, allowing the user to issue queries directly on non-integrated data, but still getting result as if the queries were issued on integrated data—it is the tool that automatically understands what data to integrate.

References

[1] S. Bergamaschi, S. Castano, M. Vincini, Semantic Integration of Semistructured and Structured Data Sources, SIGMOD Rec. 28 (1999) 54–59.

[2] S. Bergamaschi, et al., From Data Integration to Big Data Integration, in: A Comprehensive Guide Through the Italian Database Research Over the Last 25 Years, volume 31 of *Studies in Big Data*, Springer, 2018, pp. 43–59.

[3] G. Simonini, S. Bergamaschi, H. V. Jagadish, BLAST: a Loosely Schema-aware Meta-blocking Approach for Entity Resolution, PVLDB 9 (2016) 1173–1184.

[4] G. Simonini, G. Papadakis, T. Palpanas, S. Bergamaschi, Schema-agnostic Progressive Entity Resolution, in: ICDE, 2018, pp. 53–64.

[5] G. Papadakis, G. Mandilaras, L. Gagliardelli, et al., Three-dimensional Entity Resolution with JedAI, Inf. Syst. 93 (2020) 101565.

[6] G. Simonini, L. Zecchini, S. Bergamaschi, F. Naumann, Entity Resolution On-Demand, PVLDB 15 (2022) 1506–1518.

[7] L. Gagliardelli, G. Papadakis, G. Simonini, S. Bergamaschi, T. Palpanas, Generalized Supervised Meta-blocking, PVLDB 15 (2022) 1902–1910.

[8] L. Gagliardelli, L. Zecchini, D. Beneventano, et al., ECDP: A Big Data Platform for the Smart Monitoring of Local Energy Communities, in: DataPlat, volume 3135 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2022.

[9] D. Beneventano, S. Bergamaschi, L. Gagliardelli, G. Simonini, L. Zecchini, Big Data Integration & Data-Centric AI for eHealth, in: Ital-IA, 2022.

[10] L. Gagliardelli, G. Simonini, D. Beneventano, S. Bergamaschi, SparkER: Scaling Entity Resolution in Spark, in: EDBT, OpenProceedings.org, 2019, pp. 602–605.

ItaData 2022: The 1st Italian Conference on Big Data and Data Science, September 20–21, 2022, Milan, Italy

0000-0001-8087-6587 (S. Bergamaschi); 0000-0001-6616-1753 (D. Beneventano); 0000-0002-3466-509X (G. Simonini); 0000-0001-5977-1078 (L. Gagliardelli); 0000-0002-4856-0838 (L. Zecchini)

CEUR Workshop Proceedings (CEUR-WS.org)

[1] https://dbgroup.unimore.it

[2] https://www.datariver.it

[3] https://datacentricai.org