Data Integration and Querying through Scalable Neural Data Representations for Data Lakes

M2 research laboratory internship

- Start Date: February / March 2024
- Duration: 6 months
- Location: LIP6
- Co-supervisors: Rafael Angarita (rafael.angarita@lip6.fr), Hubert Naacke (hubert.naacke@lip6.fr) and Bernd Amann (bernd.amann@lip6.fr)

Context and Motivation

Data lakes are collections of massive heterogeneous datasets hosted in a variety of storage systems [7]. In contrast to data warehouses where the data has been transformed to answer specific queries, data lakes store raw unformatted data ranging from structured data such as relational tables, semi-structured data such as JSON documents, and unstructured data such as textual documents with no predefined schema or data model. Integrating such heterogeneous data is a crucial steps towards providing a unified and coherent view of the information within a data lake [2]; however, traditional integration approaches still have difficulties when dealing with disparate data and fail at uncovering hidden relations within.

Neural data representations for databases are a novel approach for revealing hidden, latent information within the data using deep learning approaches. Some applications for queries over neural representations of data include fact-checking, table metadata generation, and content prediction in relational tabular data [3, 6, 9], as well as the discovery of missing links in knowledge graphs [1]. However, neural data representations approaches cannot yet be applied to data lakes since they lack expressiveness to perform complex query and they do not handle large volumes of data efficiently.

Objective and challenges

In this project, we aim to investigate and develop new methods for integrating and querying heterogeneous data within data lakes using deep learning models. This raises the following technical challenges: how to encode the semantics of heterogeneous datasets into the embedding learning process, reconciling datasets with different schemas and with incomplete and noisy data.

Internship goals and tasks

- Literature review: Conduct a comprehensive literature review to understand existing methods and frameworks starting by the three categories presented above: Neural Tabular Data Representations, Knowledge Graph Embeddings, and Scaling Up Neural Representations of Databases.
- Data collection: Collection of a diverse range of heterogeneous data sources, including structured (e.g., tables) and unstructured data. For structured data, there exists several datasets such as WikiTables-TURL [3], WDC Web Table Corpus [5] and VizNet [4]. These datasets are used for different tasks such as question answering, semantic parsing, table retrieval, table metadata prediction and table content population. YAGO [8] is an example of a graph dataset.
- Scalable Querying of Neural Data Lakes: executing queries that necessitate the combination of results from these diverse neural data representations. This approach aims to deliver more complete answers, surpassing what can be achieved by querying each model in isolation.
- Comparative evaluation: Design experiments and benchmarks to evaluate the effectiveness of the proposed approach in generating embeddings for querying data lakes. Note that existing benchmarks are specific to certain downstream tasks such as question answering and fact checking for tabular data, and link prediction for knowledge graph; so the challenge of this tasks on designing a benchmark to test the intrinsic capabilities of neural representations of data lakes.

Required Skills

The candidate should have *excellent experience in algorithmic and programming* in Python and *advanced knowledge* in machine learning and relational and non-relational databases.

To apply, you would need to send your CV and the grades from the last three semesters of studies to the three co-supervisors (see email above).

Host team

- LIP6 Database Research Team: `http://www-bd.lip6.fr/`

References

[1] Hussein Baalbaki, Hussein Hazimeh, Hassan Harb, and Rafael Angarita. Kema++: A full representative knowledge-graph embedding model (036). *International Journal of Software Engineering and Knowledge Engineering*, 32(11n12):1619–1641, 2022.

[2] Júlia Colleoni Couto and Duncan Dubugras Ruiz. An overview about data integration in data lakes. In *2022 17th Iberian Conference on Information Systems and Technologies (CISTI)*, pages 1–7, 2022.

[3] Xiang Deng, Huan Sun, Alyssa Lees, You Wu, and Cong Yu. Turl: Table understanding through representation learning. *ACM SIGMOD Record*, 51(1):33–40, 2022.

[4] Kevin Hu, Snehalkumar'Neil'S Gaikwad, Madelon Hulsebos, Michiel A Bakker, Emanuel Zgraggen, César Hidalgo, Tim Kraska, Guoliang Li, Arvind Satyanarayan, and Çağatay Demiralp. Viznet: Towards a large-scale visualization learning and benchmarking repository. In *Proceedings of the 2019 CHI conference on human factors in computing systems*, pages 1–12, 2019.

[5] Oliver Lehmberg, Dominique Ritze, Robert Meusel, and Christian Bizer. A large public corpus of web tables containing time and context metadata. In *Proceedings of the 25th international conference companion on world wide web*, pages 75–76, 2016.

[6] Qian Liu, Bei Chen, Jiaqi Guo, Morteza Ziyadi, Zeqi Lin, Weizhu Chen, and Jian-Guang Lou. Tapex: Table pre-training via learning a neural sql executor. *arXiv preprint arXiv:2107.07653*, 2021.

[7] Fatemeh Nargesian, Erkang Zhu, Renée J Miller, Ken Q Pu, and Patricia C Arocena. Data lake management: challenges and opportunities. *Proceedings of the VLDB Endowment*, 12(12):1986–1989, 2019.

[8] Thomas Rebele, Fabian Suchanek, Johannes Hoffart, Joanna Biega, Erdal Kuzey, and Gerhard Weikum. Yago: A multilingual knowledge base from wikipedia, wordnet, and geonames. In *The Semantic Web–ISWC 2016: 15th International Semantic Web Conference, Kobe, Japan, October 17–21, 2016, Proceedings, Part II 15*, pages 177–185. Springer, 2016.

[9] Pengcheng Yin, Graham Neubig, Wen-tau Yih, and Sebastian Riedel. Tabert: Pretraining for joint understanding of textual and tabular data. *arXiv preprint arXiv:2005.08314*, 2020.