Internship topic: Data stories for interactive intentional analytics

Veronika Peralta, Patrick Marcel

November 2018

1 Context

Can data analysis be fully automated and eventually an Artificial Intelligence (AI) makes the decision? The debate around AI, especially Machine Learning (ML), and their supposed capacity at automating decision making, is very intense these days. In the database (DB) community, and more particularly in the data warehousing (DW) community, there is long tradition of having the decision maker at the center of the data analysis process. At the inverse of automated application of algorithms [4], DW has been, since its inception, all about facilitating the task of interactive exploration of a dataspace, and not let e.g., an algorithm automatically mine this space for patterns. One could even say that DW is the ancestor of the Human-In-the-Loop Data Analysis phenomenon [2].

This internship topic follows up from the reinvention of OLAP described in [9, 10], and ambitions to automatize further interactive data analysis, while letting the end user in command. This reinvention of OLAP introduces an analytics model redefining what a query is, with respect to both what users ask the system, what the answer entails, and how this answer is computed. An implementation is currently being done.

The work introduced in [10] opens several major research questions. A first question is *How to facilitate the understanding of data?* This demands to precisely define what are the answers to complex sequences of high level intentions, and package them into coherent data stories accessible to even non expert users.

As an answer to this question, authors propose that answers to intentional operators are no longer traditional sets of tuples, but dashboards including data, charts, informative summaries of KPI performance, as well as concise representations of knowledge hidden in the data.

The long term ambition is to automatically generate such dashboards based on past and current user interactions, and using data mining techniques.

2 Objectives and expected contributions

The main challenge of this internship is to define how to structure dashboards in a context where the interactive data analysis is a sequence of possibly complex queries, each being a composition of intentions, in a personalized way [8].

The detailed objectives are:

1. Study of the intentional operators proposed in [10].
2. Literature review about dashboard representation.
3. Propose a dashboard model adapted to complex intentional queries.
4. As a proof of concept, generate dashboards for a set of user explorations.

This internship may be the starting point for a PhD position, starting October 2019, with regional founding. Further challenges will be investigated during the PhD thesis, including :

- Facilitating data understanding also demands to automatically put query results in a shape that is easy to grasp and that facilitates data storytelling. The second challenge of the PhD lies in identifying the most appropriate graphical representation of query answers, and to automatically craft narratives, commenting on the highlights presented, etc. [5, 6].
- Finally, executing complex intentional statements requires an optimization phase to decide which logical operators and model mining algorithms to execute. This optimization can be thought in terms of performance, but also in terms of information content delivered and in terms of the quality of the user's experience, measuring the number of insights, controlling false discoveries, etc. [1, 7, 3, 11].

3 Application

The position is a 5 to 6 months funded master internship. Remuneration is around 600 EUR per month.

The recruited student will be supervised by Veronika Peralta and Patrick Marcel at the University of Tours, in the campus of Blois (3 place Jean-Jaurès, 41000 Blois).

Applicants are expected to be 2nd year master students in Computer Science, be skilled in databases, machine learning, programming and be fluent in English. A first experience in research is a plus.

Applicants will email the following documents to the supervisors, before February 14th, 2019:

- CV, including latest academic results,
- cover letter.

Contact:

- veronika.peralta@univ-tours.fr,
- patrick.marcel@univ-tours.fr,

References

[1] Mahfoud Djedaini, Krista Drushku, Nicolas Labroche, Patrick Marcel, Verónika Peralta, and Willeme Verdeau. Automatic assessment of interactive OLAP explorations. *To appear in Information Systems*, 2019.

[2] AnHai Doan. Human-in-the-loop data analysis: A personal perspective. In *Proceedings of the Workshop on Human-In-the-Loop Data Analytics, HILDA@SIGMOD 2018, Houston, TX, USA, June 10, 2018*, pages 1:1–1:6, 2018.

[3] Philipp Eichmann, Emanuel Zgraggen, Zheguang Zhao, Carsten Binnig, and Tim Kraska. Towards a benchmark for interactive data exploration. *IEEE Data Eng. Bull.*, 39(4):50–61, 2016.

[4] Matthias Feurer, Aaron Klein, Katharina Eggensperger, Jost Tobias Springenberg, Manuel Blum, and Frank Hutter. Efficient and robust automated machine learning. In *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pages 2962–2970, 2015.

[5] Dimitrios Gkesoulis, Panos Vassiliadis, and Petros Manousis. Cinecubes: Aiding data workers gain insights from OLAP queries. *Inf. Syst.*, 53:60–86, 2015.

[6] Jessica Hullman, Steven M. Drucker, Nathalie Henry Riche, Bongshin Lee, Danyel Fisher, and Eytan Adar. A deeper understanding of sequence in narrative visualization. *IEEE Trans. Vis. Comput. Graph.*, 19(12):2406–2415, 2013.

[7] Patrick Marcel, Nicolas Labroche, and Panos Vassiliadis. Towards a benefit-based optimizer for interactive data analysis. In *Proceedings of the 21st International Workshop on Design, Optimization, Languages and Analytical Processing of Big Data co-located with 10th EDBT/ICDT Joint Conference (EDBT/ICDT 2019), Lisboa, Portugal, March 26-29, 2019.*, 2019.

[8] Tova Milo and Amit Somech. Next-step suggestions for modern interactive data analysis platforms. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD 2018, London, UK, August 19-23, 2018*, pages 576–585, 2018.

[9] Panos Vassiliadis and Patrick Marcel. The road to highlights is paved with good intentions: Envisioning a paradigm shift in OLAP modeling. In *Proceedings of the 20th International Workshop on Design, Optimization, Languages and Analytical Processing of Big Data co-located with 10th EDBT/ICDT Joint Conference (EDBT/ICDT 2018), Vienna, Austria, March 26-29, 2018.*, 2018.

[10] Panos Vassiliadis, Patrick Marcel, and Stefano Rizzi. Beyond roll-up's and drill-down's: An intentional analytics model to reinvent OLAP. *Submitted to Information Systems*, 2019.

[11] Zheguang Zhao, Lorenzo De Stefani, Emanuel Zgraggen, Carsten Binnig, Eli Upfal, and Tim Kraska. Controlling false discoveries during interactive data exploration. In *Proceedings of the 2017 ACM International Conference on Management of Data, SIGMOD Conference 2017, Chicago, IL, USA, May 14-19, 2017*, pages 527–540, 2017.