Holistic Approach to Data Quality in Ontology-Based Data Access

Postdoctoral Researcher Position in Bordeaux, France

Keywords

Inconsistency handling, data quality, ontology-based data access, query answering, description logics, knowledge representation and reasoning

Context

Accessing the relevant information in real-world data g is difficult, time-consuming, and error-prone due to the need to integrate data across multiple heterogeneous sources. Moreover, even if this first hurdle is overcome, a perhaps even more daunting challenge arises: how to obtain reliable insights from imperfect data? It is widely acknowledged that real-world data is plagued with quality issues, such as incompleteness (missing information) and errors (false or outdated information).

The ontology-based data access (OBDA) paradigm [1,2,3] addresses the first challenge by facilitating access to (potentially heterogeneous) data sources through the use of ontologies that specify a convenient user-friendly vocabulary for query formulation (which abstracts from the way the data is stored) and capture domain knowledge that can be exploited at query time, via automated reasoning, to obtain more complete query results. While OBDA systems are growing in maturity [5], they too often fail to address the data quality issue, aside from issuing warnings when inconsistencies are discovered.

To widen the applicability of OBDA, it is essential to equip OBDA systems with appropriate mechanisms for handling imperfect data: how to obtain meaningful answers to queries posed over imperfect data, and how best to generate a high-quality version of the data? While these questions have begun to be explored [6,7], we are still far from having robust and widely applicable techniques for handling data quality in OBDA.

The postdoc position is part of the INTENDED Chair on Artificial Intelligence, whose aim is to develop intelligent, knowledge-based methods for handling imperfect data. The chair project will involve both foundational work as well as a more applied component with a hospital use case. The chair begins in September 2020 and has a duration of four years. It is funded by the French National Research Agency (ANR) and the University of Bordeaux.

Position

This is a two-year full-time postdoctoral researcher position, with an approximate monthly net salary of 2100€. The position does not include any mandatory teaching or supervisory activities, but there will be opportunities to engage in teaching and/or student supervision if desired.

The desired starting date is October 1st, 2020, but this can be (significantly) delayed if needed. The position will remain open until a suitable candidate is found.

Research Environment

The position will be based at LaBRI (Laboratoire Bordelais de Recherche en Informatique), a computer science research lab located on the University of Bordeaux's Talence campus. The lab can be easily reached from the city center of Bordeaux by tram.

The postdoctoral researcher will participate in the new research group RATIO (Reasoning with data, knowledge and constraints), which is part of a much larger Formal Methods team that brings together dozens of researchers interested in applying logical and automata-theoretic methods to a wide range of problems within computer science.

The postdoctoral researcher will collaborate with Meghyn Bienvenu (leader of the chair) and will have many opportunities to interact with the other INTENDED participants and external collaborators. The chair provides significant resources to finance trips to conferences and to host visiting researchers.

Research Topic

As part of the INTENDED project, we plan to develop a holistic approach to data quality in OBDA that incorporates existing data cleaning techniques (e.g. entity linking and statistical outliers) [8], in order to tackle a wider class of data quality issues and to improve overall results by exploiting synergies among different methods. For instance, merging distinct values using entity linking tools may both resolve violations of functionality constraints from the ontology, or bring to light additional conflicts with the ontology that would be missed otherwise, and conversely, the entity linking process could be improved by taking into account the logical constraints imposed by the ontology.

The postdoctoral researcher will participate in devising a suitable formal framework for integrating different data quality methods into the OBDA approach, and subsequently developing algorithms for computing repairs and query answering in this framework.

This is primarily a foundational research topic, but depending on the interests and aptitude of the postdoctoral researcher, it could involve a more practical component with the implementation and testing of the developed algorithms.

Profile

Applicants should hold (or be close to completing) a PhD degree in computer science and should have a strong research record in knowledge representation and reasoning or database theory (ideally, demonstrated by at least one publication in a top-tier venue).

Familiarity with logic and/or ontology languages is desirable.

Experience and interest in one or more of the following areas would be relevant and welcome: description logics, Semantic Web, automated reasoning, inconsistency handling, data integration, data cleaning, data quality, theoretical computer science.

The precise research topic may be adjusted to suit the background and interests of the postdoctoral researcher, so brilliant candidates with an interest for the overall project topic should not hesitate to apply.

Strong English language skills (reading, writing, & speaking) are expected, but knowledge of French is not required. The working language can be either French or English.

Contact

Interested candidates should contact Meghyn Bienvenu (meghyn.bienvenu@labri.fr) and provide a detailed CV (with publication list and names of 2-3 references) and a brief description of how the topic relates to their prior experience and research interests.

Bibliography

[1] A. Poggi, D. Lembo, D. Calvanese, G. De Giacomo, M. Lenzerini, and R. Rosati. Linking data to ontologies. Journal of Data Semantics, 10:133–173, 2008.

[2] G. Xiao, D. Calvanese, R. Kontchakov, D. Lembo, A. Poggi, R. Rosati, and M. Zakharyaschev. Ontology-based data access: A survey. Proceedings of the 27th International Joint Conference on Artificial Intelligence (IJCAI), pages 5511–5519, 2018.

[3] M. Bienvenu, M. Ortiz: Ontology-Mediated Query Answering with Data-Tractable Description Logics. Tutorial Notes of the 11th International Reasoning Web Summer School (LNCS 9203), pages 218-307, 2015.

[4] SNOMED CT Website: https://www.snomed.org/snomed-ct/why-snomed-ct

[5] D. Calvanese, B. Cogrel, S. Komla-Ebri, R. Kontchakov, D. Lanti, M. Rezk, M. Rodriguez-Muro, G. Xiao, "Ontop: Answering SPARQL queries over relational databases," Semantic Web Journal, vol. 8, no. 3, pp. 471-487, 2017.

[6] M. Bienvenu, C. Bourgaux: Inconsistency-Tolerant Querying of Description Logic Knowledge Bases. Tutorial Notes of the 12th International Reasoning Web Summer School (LNCS 9885), pages 156- 202, 2016.

[7] M. Bienvenu: Inconsistency Handling in Ontology-Mediated Query Answering: A Progress Report (Invited Talk). Proceedings of the 32nd International Workshop on Description Logics (DL), 2019.

[8] Z. Abedjan, X. Chu, D. Deng, R. C. Fernandez, I. Ilyas, M. Ouzzani, P. Papotti, M. Stonebraker and N. Tang, "Detecting data errors: Where are we and what needs to be done?," in VLDB, 2016.