PhD Position: Principled, Knowledge-Based Methods for Handling Imperfect Data

Keywords

Inconsistent, incomplete, & uncertain information, data quality, ontology-mediated query answering, knowledge representation & reasoning, database theory

Context

Accessing the relevant information contained in real-world data to support informed decision making 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that real-world data is plagued with quality issues, such as incompleteness (missing information) and errors (false or outdated information).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-mediated query answering (OMQA) 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. For example, querying for patients with infectious heart disease is non-trivial due to the myriad of ways such a generic condition can manifest, but by leveraging the knowledge formalized in medical ontologies (like SNOMED CT [4]), it is possible to correctly return patients diagnosed with Chagall's disease, toxoplasma myocarditis, etc.

While OMQA 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 the OMQA approach, it is essential to equip OMQA 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.

This PhD 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 is funded by the French National Research Agency (ANR) and the University of Bordeaux.

Research Topic

The PhD thesis will center on the development of principled, knowledge-based methods for handling imperfect data in the OMQA setting. Example research directions include: (1) the development of inconsistency-tolerant querying algorithms for expressive OMQA settings, (2) the integration of qualitative & quantitative reliability information for facts and constraints, and (3) the specification and analysis of inconsistency management policies.

The PhD student will carry out foundational research, involving the definition of formal frameworks for repairing and querying imperfect data, the study of the computational complexity of reasoning, and the development of reasoning algorithms. Depending on the interests and aptitude of the student, the thesis could also involve a more practical component with the implementation and testing of the developed algorithms.

Position & Research Environment

The PhD studentship is a three-year full-time position, with an approximate monthly net salary of 1700€. The position does not include any teaching obligations, but there are opportunities to engage in teaching if desired.

The PhD thesis will be co-supervised by Meghyn Bienvenu (LaBRI, Bordeaux) and Camille Bourgaux (DI ENS, Paris). The position will be based in Bordeaux in the LaBRI research lab, with regular funded stays in Paris to visit the co-supervisor.

LaBRI (Laboratoire Bordelais de Recherche en Informatique) is a computer science research lab located on the University of Bordeaux Talence campus, which can be easily reached from the city center of Bordeaux by tram. The PhD student will participate in the new research group RATIO (Reasoning with data, knowledge and constraints).

Candidate Profile

At the start of the PhD, the candidate must hold a Master's degree in computer science (or possibly mathematics, if accompanied by relevant computer science experience).

This position is is best suited to candidates who have prior experience with knowledge representation and reasoning (especially: description logics, non-monotonic reasoning), database theory, or Semantic Web (ontologies).

Candidates must demonstrate familiarity with propositional and first-order logic and basic notions of computational complexity.

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.

How to Apply

The PhD position is currently available and will remain open until a suitable candidate is found. We welcome applications from students who are currently attending a Master's program, and we can propose a Master's internship on a related topic.

Potential candidates, either for the PhD position or for a related Master's internship, should contact the two supervisors by email:

- Meghyn Bienvenu (meghyn.bienvenu@labri.fr)
- Camille Bourgaux (camille.bourgaux@ens.fr)

The email should include a CV, university transcripts, and a short 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. In 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: A Short Survey on Inconsistency Handling in Ontology-Mediated Query Answering. Künstliche Intelligenz 34(4): 443-451, 2020.