Internship topic

Query answer diversity in ontology-based data access

Where: ENS-PSL, Paris, or LaBRI, Bordeaux

With whom: Michaël Thomazo (ENS-PSL, Inria) and Meghyn Bienvenu (CNRS & Université de Bordeaux)

When: 5/6 months, starting from February 2026 (flexible)

Funding: "Gratification". Possible follow-up PhD funding (at ENS).

Setting and Motivation

A common issue when querying data is that the set of query answers is too large to be examined, let alone understood, by a human, making it more reasonable to present users with a small subset of the answers. It is typically desirable to select a diverse set of query answers, which provide a representative summary of the answer set, rather than choosing a random or arbitrary set, which might contain several highly similar query results. This is especially important when query results are used to support decision making (e.g. one would like to see a variety of different travel itineraries or product models among those satisfying the queried criteria) or data exploration (e.g. once a user has viewed a small, initial set of results, they may request further tuples that are (dis)similar to particular output tuples). Query answer diversity has been explored, both from practical and theoretical angles, by researchers in the database community and remains an active research topic [1,2,3]. Similar notions have also been considered in other areas, such as information retrieval and reasoning systems (SAT, ASP) [4].

Ontology-based data access (OBDA [5], see [6,7] for introductions) is a paradigm that enables users to query heterogeneous data sources through a conceptual model, also known as ontology, without needing to know the details of the underlying data sources. In addition to providing a high-level unified vocabulary for query formulation, the ontology captures domain knowledge, which is exploited by reasoning algorithms to compute the set of all certain query answers (i.e. those that can be inferred by combining information from the data sources and ontology). Although the issue of large result sets arises also in the OBDA setting (especially since the ontology can lead to many additional answers being inferred), the problem of providing diverse answers has not been studied so far in the context of OBDA, and we aim at filling this gap with this internship.

Internship Topic and Goals

The first step in the internship will be to gain familiarity with existing approaches for query answer diversity that have been proposed in databases and to adapt them in the simplest possible way to fit the OBDA setting. In a second step, the intern will explore how to define novel notions of diversity specifically targeting the OBDA setting, which will take the ontology axioms and reasoning steps into account. The internship (and the possible follow-up PhD thesis) will aim to make conceptual, algorithmic, and complexity-theoretic contributions: proposing and comparing new formal definitions for query answer diversity in OBDA, devising or adapting algorithms to compute diverse query answers, and analyzing the complexity of the associated computational tasks to identify tractable cases and sources of intractability.

If both the student and supervisors are happy with the collaboration, a PhD funding (at ENS Ulm, Paris) is available to further the study.

Pre-requisites

Familiarity and interest in first-order logic, notions of complexity theory. Prior exposure to knowledge representation formalisms, automated reasoning, and/or database theory would be a plus, but is not mandatory.

How to apply

Contact Meghyn Bienvenu (first.last at u-bordeaux.fr) and Michaël Thomazo (first.last at inria.fr), with:

— a CV;

— the list of courses followed that have a link with the internship topic;

— a transcript of grades available so far.

References

[1] T. Deng, W. Fan. On the complexity of query result diversification. *ACM Transactions on Database Systems (TODS)*, 2014.

[2] K. Zheng, H. Wang, Z. Qi, J. Li, and H. Gao. A survey of query result diversification. *Knowledge and Information Systems*, 2017.

[3] T. C. Merkl, R. Pichler, and S. Skritek. Diversity of answers to conjunctive queries. *Logical Methods in Computer Science (LMCS)*, 2025.

[4] T. Eiter, E. Erdem, H. Erdogan, M. Fink. Finding similar/diverse solutions in answer set programming. *Theory and Practice of Logic Programming (TPLP)*, 2013.

[5] A. Poggi, D. Lembo, D. Calvanese, G. De Giacomo, M. Lenzerini, R. Rosati. Linking data to ontologies. *Journal of Data Semantics*, 2008.

[6] M. Bienvenu, M. Ortiz. Ontology-mediated query answering with data-tractable description logics. *Reasoning Web*, 2015.

[7] G. Xiao, D. Calvanese, R. Kontchakov, D. Lembo, A. Poggi, R. Rosati, M. Zakharyaschev. Ontology-based data access: A survey. In *Proceedings of IJCAI*, 2018.

Sujet de stage

Réponses diverses à des requêtes en présence d'ontologies

Où : DIENS, Paris ou LaBRI, Bordeaux

Avec qui : Michaël Thomazo (ENS-PSL, Inria) et Meghyn Bienvenu (CNRS & Université de Bordeaux)

Quand : 5/6 mois, début à partir de février 2026

Financement : « Gratification ». Poursuite en thèse possible (à l'ENS).

Contexte et motivation

Un problème courant lors de l'interrogation de données est que l'ensemble des réponses aux requêtes est trop volumineux pour être examiné, et encore moins compris, par un humain. Cela motive de présenter aux utilisateurs un sous-ensemble limité des réponses. On souhaite alors sélectionner un ensemble diversifié de réponses aux requêtes, qui fournissent une synthèse représentative de l'ensemble des réponses, plutôt que de choisir un ensemble aléatoire ou arbitraire, qui pourrait contenir plusieurs résultats très similaires. Cela est particulièrement important lorsqu'on utilise les résultats des requêtes pour appuyer la prise de décision (par exemple, on souhaiterait voir une variété d'itinéraires de voyage ou de modèles de produits parmi ceux satisfaisant aux critères interrogés) ou lors de l'exploration de données (par exemple, un petit ensemble initial de résultats ayant été visualisé, obtenir des tuples supplémentaires (dis)similaires à certains tuples de sortie spécifiques). La diversité des réponses aux requêtes a été explorée, à la fois d'un point de vue pratique et théorique, par les chercheurs de la communauté des bases de données et reste un sujet de recherche actif [1,2,3]. Des notions similaires ont également été envisagées dans d'autres domaines, tels que la recherche d'information et les systèmes de raisonnement (SAT, ASP) [4].

L'accès aux données au travers d'une ontologie (OBDA [5], voir [6,7] pour des introductions) est un paradigme qui permet aux utilisateurs de requêter des sources de données hétérogènes via un modèle conceptuel, également appelé ontologie, sans avoir besoin de connaître les détails des sources de données sous-jacentes. En plus de fournir un vocabulaire unifié de haut niveau pour la formulation des requêtes, l'ontologie capture des connaissances du domaine, qui sont exploitées par les algorithmes de raisonnement pour calculer l'ensemble de toutes les réponses certaines aux requêtes (c'est-à-dire celles qui peuvent être inférées en combinant l'information des sources de données et de l'ontologie). Bien que le problème des ensembles de résultats volumineux se produise également dans le cadre OBDA (surtout puisque l'ontologie peut entraîner l'inférence de nombreux résultats supplémentaires), le problème de fournir des réponses diversifiées n'a pas été étudié jusqu'à présent dans le contexte de l'OBDA, et nous visons à combler ce vide avec ce stage.

Sujet et objectifs du stage

La première étape du stage sera de se familiariser avec les approches existantes pour la diversité des réponses aux requêtes proposées dans les bases de données et d'adapter ces approches de la manière la plus simple possible au cadre OBDA. Dans une deuxième étape, le ou la stagiaire explorera comment définir de nouvelles notions de diversité spécifiquement ciblées sur le cadre OBDA, qui prendront en compte les axiomes de l'ontologie et les étapes de raisonnement. Le stage (et la thèse de doctorat possible qui en découlerait) vise à apporter des contributions conceptuelles, algorithmiques et théoriques : proposer et comparer de nouvelles définitions formelles de la diversité des réponses aux requêtes dans l'OBDA, concevoir ou adapter des algorithmes pour calculer des

réponses aux requêtes diversifiées, et analyser la complexité des tâches computationnelles associées afin d'identifier les cas tractables et les sources d'intractabilité.

Si tous les membres de la collaboration sont satisfaits, un financement de thèse (à l'ENS-PSL, Paris) est disponible pour poursuivre l'étude.

Prérequis

Familiarité et intérêt pour la logique du premier ordre, les notions de théorie de la complexité. Une exposition antérieure aux formalismes de représentation des connaissances, au raisonnement automatisé et/ou à la théorie des bases de données serait un atout, mais n'est pas obligatoire.

Comment postuler

Contactez Meghyn Bienvenu (first.last @ u-bordeaux.fr) et Michaël Thomazo (first.last @ inria.fr), en joignant :

— un CV ;

— la liste des cours suivis ayant un lien avec le sujet du stage ;

— un relevé de notes disponible jusqu'à présent.

Références

[1] T. Deng, W. Fan. On the complexity of query result diversification. *ACM Transactions on Database Systems (TODS)*, 2014.

[2] K. Zheng, H. Wang, Z. Qi, J. Li, and H. Gao. A survey of query result diversification. *Knowledge and Information Systems*, 2017.

[3] T. C. Merkl, R. Pichler, and S. Skritek. Diversity of answers to conjunctive queries. *Logical Methods in Computer Science (LMCS)*, 2025.

[4] T. Eiter, E. Erdem, H. Erdogan, M. Fink. Finding similar/diverse solutions in answer set programming. *Theory and Practice of Logic Programming (TPLP)*, 2013.

[5] A. Poggi, D. Lembo, D. Calvanese, G. De Giacomo, M. Lenzerini, R. Rosati. Linking data to ontologies. *Journal of Data Semantics*, 2008.

[6] M. Bienvenu, M. Ortiz. Ontology-mediated query answering with data-tractable description logics. *Reasoning Web*, 2015.

[7] G. Xiao, D. Calvanese, R. Kontchakov, D. Lembo, A. Poggi, R. Rosati, M. Zakharyaschev. Ontology-based data access : A survey. In *Proceedings of IJCAI*, 2018.