Towards an Equivalence Degree of $\mathcal{EL}$ CQs* (Extended abstract)

Oliver Fernández Gil and Anni-Yasmin Turhan

Theoretical Computer Science, TU Dresden, Germany
firstname.lastname@tu-dresden.de

We report on initial steps towards designing similarity measures for conjunctive queries formulated over $\mathcal{EL}$ ontologies. In applications that use conjunctive queries (CQs) it is often the question how similar two queries are to each other. We consider here similarity of CQs in general and develop a particular family of conjunctive query similarity measures (CQSMs).

Assessing the similarity in the presence of TBoxes has been considered for *concept similarity measures* (CSMs), which are functions that compute for a pair of concepts a value from the unit interval. As similarity per se is not a formal notion and similarity measures can be defined in an arbitrary fashion, often a catalog of properties is devised that a "well-behaved" measure should fulfill [3]. The overall idea is to perceive similarity of concepts as a form of gradual equivalence. CSMs with these properties have been devised for $\mathcal{EL}$-concepts in [3], albeit only for acyclic TBoxes, and in [2,5] for general TBoxes. These CSMs implement gradual equivalence by computing a combination of the mutual subsumption degree of the two concepts. More precisely, this is done by finding simulations between the canonical models of the TBox and the respective concepts— as such simulations characterize subsumption in $\mathcal{EL}$ [4].

We want to transfer this approach to CQs and perceive similarity of CQs as a form of "gradual query equivalence" which yields measures that are defined independent of the data. However, the transfer is not straightforward for several reasons. While the existence of simulations characterize equivalence and subsumption of $\mathcal{EL}$ concepts w.r.t. a TBox, CQs require homomorphisms and the compact canonical model representation used for concepts does no longer suffice. In addition, pairs of CQs can differ in their arity or in the set of individual names occurring in them. In this initial investigation we concentrate on CQSMs for CQs that are rooted (every quantified variable can be reached from an answer variable) and that are *uniform* in the sense that the input queries must have the same number of answer variables and the same set of individuals. For the empty TBox, it is well-known that equivalence of queries can be characterized by the existence of isomorphism between the queries, provided that the queries are minimized [1]. Hence, we apply techniques for graph similarity to define a similarity measure for conjunctive queries, more precisely we use error-tolerant graph matchings.

* Supported by DFG in the RTG 1907 (RoSI) and grant BA 1122/20-1.

In the presence of a TBox, however, the measure has to take the TBox information into account, which means that equivalence of CQs cannot be tested by a comparison of the plain query graphs.

Our approach to develop a uniform CQSM for rooted $\mathcal{EL}$ CQs consists of several steps. First, we define a CQSM for plain query graphs in terms of graph edit distance, where the costs of some edit operations are adapted to the specific setting of CQs. So, for instance, to compare two nodes with concept labels, the similarity of these node labels, is assessed by the CSM from [2].

The second step is to extend this measure to consider non-empty $\mathcal{EL}$ TBoxes. To this end, we employ the result that CQ equivalence w.r.t. an $\mathcal{EL}$ TBox can be reduced to the case with an empty TBox, by rewriting the queries with TBox information. The general idea is to rewrite the input CQs and the TBox, by using essentially the "rolling-up technique" [6], which replaces in the input queries certain tree structures with fresh concept names. The TBox is then extended with concept definitions for the fresh concept names that capture the trees pruned from the two queries. Next, the pruned queries are rewritten again by adding new atoms that capture the consequences of the extended TBox for the query structure. One can show that the final CQs are equivalent w.r.t. the empty TBox iff the initial ones are equivalent w.r.t. the original TBox. Thus, the similarity between the input queries is the similarity between the rewritten queries. By using the CSM from [2], the comparison of the nodes for individuals (without their relational neighborhood) still considers the information in the original TBox.

The contributions are the following: we define CQSMs and formalize a collection of useful properties for CQSMs that can ensure predictable behavior of the measure to a certain extent. Furthermore, we construct a (family of) CQSMs in $\mathcal{EL}$ that fulfill some of the most important properties by following the approach just described. We show that the resulting CQSMs fulfill equivalence invariance and equivalence closure.

References

1. Ashok K. Chandra and Philip M. Merlin. Optimal implementation of conjunctive queries in relational data bases. In *Proc. of the 9th ACM Symp. on Theory of Computing (STOC'77)*, pages 77–90. ACM, 1977.
2. Andreas Ecke, Rafael Peñaloza, and Anni-Yasmin Turhan. Similarity-based relaxed instance queries. *J. Applied Logic*, 13(4):480–508, 2015.
3. Karsten Lehmann and Anni-Yasmin Turhan. A framework for semantic-based similarity measures for $\mathcal{ELH}$-concepts. In *Proc. of the 13th Eur. Conf. on Logics in Artificial Intelligence (JELIA'2012)*, volume 7519 of *LNCS*, pages 307–319. Springer, 2012.
4. Carsten Lutz and Frank Wolter. Conservative extensions in the lightweight description logic $\mathcal{EL}$. In *Proc. of the 21st Int. Conf. on Automated Deduction (CADE 2007)*, volume 4603 of *LNCS*, pages 84–99. Springer, 2007.
5. Teeradaj Racharak, Boontawee Suntisrivaraporn, and Satoshi Tojo. Personalizing a concept similarity measure in the description logic $\mathcal{ELH}$ with preference profile. *Computing and Informatics*, 37(3):581–613, 2018.

6. Riccardo Rosati. On conjunctive query answering in $\mathcal{EL}$. In *Proc. of the 2007 Description Logic Workshop (DL 2007)*, volume 250 of *CEUR*, 2007.