A Theoretical View on Reverse Engineering Problems for Database Query Languages*

Pablo Barceló[1][0000−0003−2293−2653]

DCC, University of Chile & IMFD Chile
pbarcelo@dcc.uchile.cl

Abstract. A typical reverse engineering problem for a query language L is as follows: Given a database D and two sets P and N of tuples over D labeled as *positive* and *negative* examples, respectively, is there a query q in L that *explains* P and N, i.e., the evaluation of q on D contains all positive examples in P and none of the negative examples in N? Applications of reverse engineering problems include database explanations, data exploration, data security, relational classifier engineering, and the study of the expressiveness of query languages.
In this talk I will present a family of tests that solve the reverse engineering problem described above for several query languages of interest, e.g., FO, CQ, UCQs, RPQs, CRPQs, etc. We will see that in many cases such tests directly provide optimal bounds for the problem, as well as for the size of the smallest query that explains the given labeled examples. I will also present restrictions that alleviate the complexity of the problem when it is too high. Finally, I will develop the relationship between reverse engineering and a separability problem recently introduced in the database theory literature to assist the task of relational classifier engineering with data management tools.

Keywords: reverse engineering · definability · query by example · separability.

1 Bibliographical Review

Good introductions to the applications of reverse engineering problems in data exploratory analysis can be found in [1, 13, 12]. The study of theoretical issues related to reverse engineering problems has received quite some attention in different contexts; e.g., for first-order logic and the class of conjunctive queries over relational databases [19, 16, 11, 7, 3, 18, 15, 17, 10]; for regular path queries over graph databases [2, 6]; for SPARQL queries over RDF [4]; for tree patterns over XML [8, 14]; and for description logics in the setting of ontology-based data access [9]. The application of reverse engineering methods to relational classifier engineering can be found in [5].

* The author is supported by the Millennium Institute for Foundational Research on Data and Fondecyt grant 1170109.

References

1. Ziawasch Abedjan, Lukasz Golab, and Felix Naumann. Data profiling: A tutorial. In *SIGMOD*, pages 1747–1751, 2017.
2. Timos Antonopoulos, Frank Neven, and Frédéric Servais. Definability problems for graph query languages. In *ICDT*, pages 141–152, 2013.
3. Marcelo Arenas and Gonzalo I. Díaz. The exact complexity of the first-order logic definability problem. *ACM Trans. Database Syst.*, 41(2), to 2016.
4. Marcelo Arenas, Gonzalo I. Díaz, and Egor V. Kostylev. Reverse engineering sparql queries. In *WWW*, 2016.
5. Pablo Barceló, Alex Baumgartner, Victor Dalmau, and Benny Kimelfeld. Regularizing conjunctive features for classification. In *PODS*, 2019.
6. Angela Bonifati, Radu Ciucanu, and Aurélien Lemay. Learning path queries on graph databases. In *EDBT*, pages 109–120, 2015.
7. Angela Bonifati, Radu Ciucanu, and Slawek Staworko. Learning join queries from user examples. *ACM Trans. Database Syst.*, 40(4):24, 2016.
8. Sara Cohen and Yaacov Y. Weiss. Learning tree patterns from example graphs. In *ICDT*, pages 127–143, 2015.
9. Víctor Gutiérrez-Basulto, Jean Christoph Jung, and Leif Sabellek. Reverse engineering queries in ontology-enriched systems: The case of expressive horn description logic ontologies. In *IJCAI*, pages 1847–1853, 2018.
10. Dmitri V. Kalashnikov, Laks V. S. Lakshmanan, and Divesh Srivastava. Fastqre: Fast query reverse engineering. In *SIGMOD*, pages 337–350, 2018.
11. Hao Li, Chee-Yong Chan, and David Maier. Query from examples: An iterative, data-driven approach to query construction. *PVLDB*, 8(13):2158–2169, 2015.
12. Denis Mayr Lima Martins. Reverse engineering database queries from examples: State-of-the-art, challenges, and research opportunities. *Information Systems*, 83:89 – 100, 2019.
13. Davide Mottin, Matteo Lissandrini, Yannis Velegrakis, and Themis Palpanas. New trends on exploratory methods for data analytics. *PVLDB*, 10(12):1977–1980, 2017.
14. Slawek Staworko and Piotr Wieczorek. Characterizing XML twig queries with examples. In *ICDT*, pages 144–160, 2015.
15. Balder ten Cate and Víctor Dalmau. The product homomorphism problem and applications. In *ICDT*, pages 161–176, 2015.
16. Quoc Trung Tran, Chee Yong Chan, and Srinivasan Parthasarathy. Query reverse engineering. *VLDB J.*, 23(5):721–746, 2014.
17. Yaacov Y. Weiss and Sara Cohen. Reverse engineering spj-queries from examples. In *PODS*, pages 151–166, 2017.
18. Ross Willard. Testing expressibility is hard. In *CP*, pages 9–23, 2010.
19. Meihui Zhang, Hazem Elmeleegy, Cecilia M. Procopiuc, and Divesh Srivastava. Reverse engineering complex join queries. In *SIGMOD*, pages 809–820, 2013.