University of Münster

Department of Information Systems

Not Only SQL: Efficient Positive and Unlabeled Machine Learning for Query By Example

Master Thesis

submitted by

-

Chair of Data Science:
Machine Learning and Data Engineering

Principal Supervisor	Prof. Dr. Fabian Gieseke
Supervisor	Dr. Denis Martins
	Chair for Data Science:
	Machine Learning and
	Data Engineering
Student Candidate	-
Matriculation Number	-
Field of Study	Information Systems
Contact Details	student@uni-muenster.de
Submission Date	13.09.2021

Not Only SQL: Efficient Positive and Unlabeled Machine Learning for Query By Example

Formulating database queries in terms of SQL is often a challenge for a growing number of non-database experts (e.g., biologists, journalists, business administrators) that are required to access and explore data. *Query By Example* (QBE) methods [Zlo75] offer an alternative mechanism where users can retrieve information from large databases using *data examples* that characterize their intent without having to write complex SQL queries.

QBE is commonly defined as the problem of finding a query Q over a database D and a set of data examples E such that $E \subseteq Q(D)$. Figure 1 illustrates a QBE application where a non-technical user wants to find American sportive cars in a large car database.

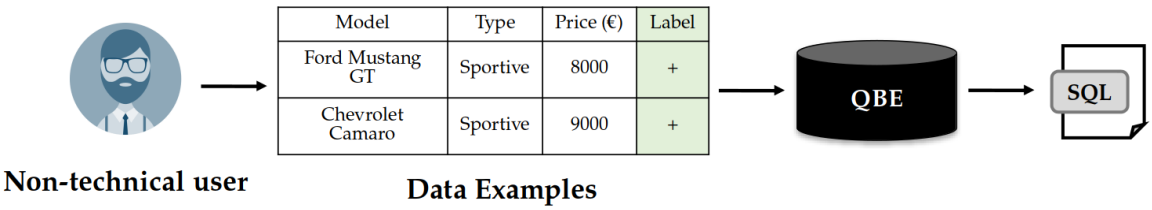

Figure 1 Query By Example in relational databases.

Traditional QBE methods such as SQLSynthesizer [ZS13] and TALOS [TCP14] address QBE under the setting of a classification problem where machine learning models are trained to classify data objects (i.e., tuples) in the database as positive or negative, whenever they match the data examples or not, respectively. After training, the model is transformed into a series of query predicates and applied to the complete database. However, both methods require data examples to fully characterize the user intention [FM19]. That is, they require a fully labeled training dataset. In practical QBE applications, users often provide only a *small subset of the positive data class*.

This thesis aims to explore efficient *positive and unlabeled learning* (PUL) techniques [BD20] for QBE over large databases. Recent PUL techniques [KTH19] fit the unique setting of QBE and are interesting alternatives to existing methods.

Bibliography

[BD20] Jessa Bekker and Jesse Davis. "Learning from positive and unlabeled data: A survey". In: *Machine Learning* 109.4 (2020), pp. 719–760.

[FM19] Anna Fariha and Alexandra Meliou. "Example-Driven Query Intent Discovery: Abductive Reasoning Using Semantic Similarity". In: *Proc. VLDB Endow.* 12.11 (July 2019), pp. 1262–1275. ISSN: 2150-8097. DOI: 10.14778/3342263.3342266. URL: https://doi.org/10.14778/3342263.3342266.

[KTH19] Masahiro Kato, Takeshi Teshima, and Junya Honda. "Learning from Positive and Unlabeled Data with a Selection Bias". In: *International Conference on Learning Representations*. 2019. URL: https://openreview.net/forum?id=rJzLciCqKm.

[TCP14] Quoc Trung Tran, Chee-Yong Chan, and Srinivasan Parthasarathy. "Query reverse engineering". In: *The VLDB Journal* 23.5 (2014), pp. 721–746.

[Zlo75] Moshé M. Zloof. "Query-by-Example: the Invocation and Definition of Tables and Forms". In: *Proceedings of the International Conference on Very Large Data Bases, September 22-24, 1975, Framingham, Massachusetts, USA*. 1975, pp. 1–24.

[ZS13] Sai Zhang and Yuyin Sun. "Automatically Synthesizing SQL Queries from Input-output Examples". In: *Proceedings of the 28th IEEE/ACM International Conference on Automated Software Engineering*. ASE'13. Silicon Valley, CA, USA: IEEE Press, 2013, pp. 224–234. ISBN: 978-1-4799-0215-6. DOI: 10.1109/ASE.2013.6693082.