Technology-Enabled Database Education: Challenges and Opportunities

Hui Li
Xidian University
hli@xidian.edu.cn

Motivation: *User-friendly* systems and tools are paramount for facilitating learning of database systems. In many universities around the world, database systems courses are supplemented with the use of industrial-strength relational database management systems (RDBMS). Although there exist a few efforts on mini RDBMS for learners to practice implementing a database kernel, *e.g.,* Minibase [6], unfortunately, there is a lack of learner-friendly technological support to facilitate the learning of various core components of a database systems course, such as database storage and indexing, relational query processing and optimization, and transaction management.

Key challenges: Specifically, our longitudinal study and observation of learners learning the topic of relational query processing reveal that they face challenges in understanding the query execution plan (QEP) of an SQL query, understanding the alternative plan choices made by the underlying query optimizer, estimating cost of a plan, etc [7, 16]. We advocate that technologies can be leveraged to improve the understanding of the above issues. For instance, since natural language (NL)-based narratives aided with visual examples (as in textbooks and lectures) have been the traditional mode of learning for decades, an intuitive NL-based description of a QEP can greatly augment learning of the execution strategies of SQL queries by an RDBMS, where recent achievements in natural language processing may suggest potential solutions.

Our efforts and experiences: To address the aforementioned challenges of learning relational query processing, we are building technological frameworks to effectively supplement the learning of various subtopics. Our recent efforts in these aspects have demonstrated positive effects in practical educational scenarios.

NEURON [3, 13] is the first system that takes advantage of a rule-based interpretation engine to generate the NL description for QEP in PostgreSQL. Specifically, NEURON first parses and transforms a QEP into an operator tree, and then traverses the tree to generate a NL description of the node based on rule-based NL templates and information encoded in the QEP. LANTERN [1, 9, 18] further makes the solution generalizable and psychology-aware, *i.e.,* generate diversified descriptions for the same QEP to avoid repetition of output messages that can lead to annoyance and boredom [8]. MOCHA [2, 17] aids learner-friendly interaction and visualization of the impact of alternative physical operator choices on a selected QEP for a given SQL query. ARENA [4, 19] further presents a novel problem called the *informative plan selection problem* (TIPS) which aims to discover a set of top-k *informative* alternative query plans from the underlying plan space. DBinsight [15] allows learners to visualize and interact with the optimization pipeline, *e.g.,* to turn on / off specific optimization rules, and manually adjust the estimated cardinality, *etc.*, so that it is easy for learners to observe and understand how each individual optimization rule (*resp.,* cardinality) can affect the choice of plans (*resp.,* a physical operator).

All these tools are publicly-available for pedagogical use [5]. Practical experiences and academic outcomes of learners taking database systems course in both Xidian University and Nanyang Technological University demonstrate that these platforms seem to facilitate understanding of various subtopics of relational query processing as well as improve academic outcomes on average. We also observe that these tools have been accessed by 74 universities and database companies around the world. More than 1300 users have used these tools.

Opportunities: We believe that these tools pave the way for designing technology-enabled solutions to improve learning and understanding of various complex topics of relational query processing. There are plenty of novel and non-trivial research challenges still waiting to be overcome, *e.g.,* how the cost of a physical operator/plan is estimated. Moreover, it is easy to see that technology-enabled learning can be extended to other database education topics [11], such as enabling technologies to facilitate the learning of SQL [10, 12, 14], developing user-friendly tools to understand different mechanisms for concurrency control, exploring opportunities to unveil the recovery procedure with the help of logs, and leveraging LLM in database education.

REFERENCES

[1] LANTERN software. `https://dbedu.xidian.edu.cn/lantern/`.

[2] MOCHA software. `https://dbedu.xidian.edu.cn/mocha/`.

[3] NEURON software. `https://dbedu.xidian.edu.cn/neuron/`.

[4] ARENA software. `https://dbedu.xidian.edu.cn/arena/`.

[5] DBEDU platform. `https://dbedu.xidian.edu.cn/`.

[6] Minibase project. `https://pages.cs.wisc.edu/~dbbook/openAccess/Minibase/minibase.html`.

[7] S. S. Bhowmick. Human Learners of Relational Query Processing: Who Cares? *In DataEd@SIGMOD*, 2023.

[8] J. T. Cacioppo and R. E. Petty. Effects of Message Repetition and Position on Cognitive Response, Recall, and Persuasion. *Journal of Personality and Social Psychology*, 37(1), 1979.

[9] P. Chen, H. Li, S. S. Bhowmick, S. R. Joty, W. Wang. LANTERN: Boredom-conscious Natural Language Description Generation of Query Execution Plans for Database Education. *In SIGMOD*, 2022.

[10] Y. Hu, Z. Miao, Z. Leong, H. Lim, Z. Zheng, S. Roy, K. Stephens-Martinez, J. Yang. I-Rex: An Interactive Relational Query Debugger for SQL. *In ACM Technical Symposium on Computer Science Education (SIGCSE)*, 2022.

[11] Z. Ives, J. Gehrke, J. Giceva, A. Kumar, R. Pottinger. VLDB Panel Summary: "The Future of Data(base) Education: Is the Cow Book Dead?". *SIGMOD Rec.*, 50(3), 2021.

[12] A. Leventidis, J. Zhang, C. Dunne, W. Gatterbauer, H. V. Jagadish, M. Riedewald. QueryVis: Logic-based Diagrams help Users Understand Complicated SQL Queries Faster. *In SIGMOD*, 2020.

[13] S. Liu, S. S. Bhowmick, W. Zhang, S. Wang, W. Huang, S. Joty. NEURON: Query Optimization Meets Natural Language Processing For Augmenting Database Education. *In SIGMOD*, 2019.

[14] Z. Miao, S. Roy, J. Yang. Explaining Wrong Queries Using Small Examples. *In SIGMOD*, 2019.

[15] Y. Rong, H. Li, K. Zhao, X. Gao, J. Cui. DBinsight: A Tool for Interactively Understanding the Query Processing Pipeline in RDBMSs. *In CIKM*, 2022.

[16] S. S Bhowmick and H. Li. Towards Technology-Enabled Learning of Relational Query Processing. In *IEEE Data Eng. Bull.* 45(3), 2022.

[17] J. Tan, D. Yeo, R. Neoh, H.-E. Chua, S. S. Bhowmick. MOCHA: A Tool for Visualizing Impact of Operator Choices in Query Execution Plans for Database Education. *PVLDB*, 15(12), 2022.

[18] W. Wang, S. S. Bhowmick, H. Li, S. Joty, S. Liu, P. Chen. Towards Enhancing Database Education: Natural Language Generation Meets Query Execution Plans. *In SIGMOD*, 2021.

[19] H. Wang, H. Li, S. S. Bhowmick, B. Xu. ARENA: Alternative Relational Query Plan Exploration for Database Education. *In SIGMOD*, 2023.