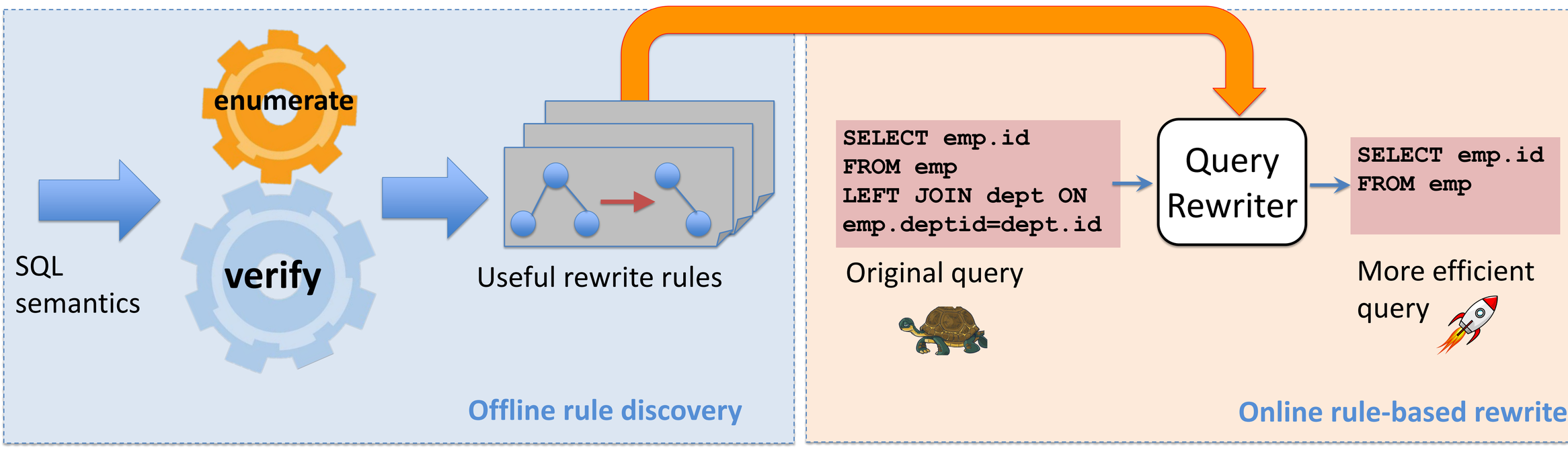
FMitF Track I

Automatic Discovery and Verification of Database Query Transformations

Jinyang Li (New York University) Ruzica Piskac (Yale University)

https://news.cs.nyu.edu/~jinyang/awards/nsf-fmitf-2220407


Goal: find new SQL rewrite rules and use them to improve query performance



Challenges

Rule discovery:

- How to define the search space of rewrite rules?
  - A rewrite rule consists of a generic source and destination query pair. How to enumerate rules?

Rule verification (<=> Equivalence verification)

- Prior work on SQL equivalence checking (UDP, SPES) is limited.
  - They translate query Q to algebraic formula f, and determine $f_1 = f_2$? by normalizing f_1/f_2 and comparing syntax structures

Scientific Impact

- The first tool to enable automatic discovery of new SQL rewrite rules.
  - Using discovered new rules, we can optimize 247 queries
- A new state-of-the SQL equivalence checker (SQLSolver) based on the theory of LIA*
  - LIA* can translate a query's algebraic expression to first order logic in a principled way
- A new theory that extends general SMT theory with multisets and Presburger arithmetic

Solution (Rule Discovery)

- Model a rewrite rule as a triplet $<q_s, q_d, c>$

Templatized query with symbols for table and column names

Constraints that relate different symbols

- Enumerate pairs of query templates q_s, q_d and search for constraints c s.t. c → $q_s \equiv q_d$ through successive relaxation.
- To verify a rule, we translate the correctness condition, c → $q_s \equiv q_d$, into FOL to be checked by an SMT solver.

Solution (Rule Verfication)

- Translate query from SQL to extended U-expression.
  - Q is expressed as a function f(t) that returns the multiplicity of any tuple t according to Q's output.
  - To prove $Q_1=Q_2$, we prove $\forall t.(f_1(t)=f_2(t))$

sum w/ unbounded domain!

- Challenge: Select x from R → $f(t) = \sum_{t'} ite(t'.x=t,1,0)$
- SQLSolver is based on LIA* [Piskac and Kuncak CAV'08]
  - LIA* can reason about multisets with unbounded sizes and model unbounded sum of integers
  - U-expr → LIA* → LIA

Solve using SMT solver

Broader Impact:

Publications:

- WeTune: Automatic Discovery and Verification of Query Rewrite Rules, SIGMOD 2022
- Proving Query Equivalence Using Linear Integer Arithmetic, SIGMOD 2023
- Automated validating and Fixing of Text-to-SQL Transaction with Execution Consistency, under submission

Open-source software:

SQLSolver: https://sqlsolver.systems/sqlsolver

Educational Activities and Outreach

- Developed graduate-level course "machine learning systems"
  - Relevant course component: verifying dataflow graph transformation in ML systems.
- GSTEM (summer research for high school students) and Pathways to AI (summer research for undergraduates)