WeTune: Automatic Discovery and Verification of Query Rewrite Rules

Zhaoguo Wang[1,2], Zhou Zhou[1,2], Yicun Yang[1,2], Haoran Ding[1,2]
Gansen Hu[1,2], Ding Ding[3], Chuzhe Tang[1,2], Haibo Chen[1,2], Jinyang Li[3]
[1]Institute of Parallel and Distributed Systems, Shanghai Jiao Tong University
[2]Engineering Research Center for Domain-specific Operating Systems, Ministry of Education, China
[3]Department of Computer Science, New York University

Abstract

Query rewriting transforms a relational database query into an equivalent but more efficient one, which is crucial for the performance of database-backed applications. Such rewriting relies on pre-specified rewrite rules. In existing systems, these rewrite rules are discovered through manual insights and accumulate slowly over the years.

In this paper, we present WeTune, a rule generator that automatically discovers new rewrite rules. Inspired by compiler superoptimization, WeTune enumerates all valid logical query plans up to a certain size and tries to discover equivalent plans that could potentially lead to more efficient rewrites. The core challenge is to determine which set of conditions (aka constraints) allows one to prove the equivalence between a pair of query plans. We address this challenge by enumerating combinations of "interesting" constraints that relate tables and their attributes between each pair of queries. We also propose a new SMT-based verifier to verify the equivalence of a query pair under different enumerated constraints. To evaluate the usefulness of rewrite rules discovered by WeTune, we apply them on the SQL queries collected from the 20 most popular open-source web applications on GitHub. WeTune successfully optimizes 247 queries that existing databases cannot optimize, resulting in substantial performance improvements.

CCS Concepts

• **Information systems** → **Query optimization**; • **Theory of computation** → **Program verification**.

Keywords

Query Rewriting, Rewrite Rule Discovery, SQL Solver

ACM Reference Format:
Zhaoguo Wang, Zhou Zhou, Yicun Yang, Haoran Ding, Gansen Hu, Ding Ding, Chuzhe Tang, Haibo Chen, and Jinyang Li. 2022. WeTune: Automatic Discovery and Verification of Query Rewrite Rules. In *Proceedings of the 2022 International Conference on Management of Data (SIGMOD '22), June 12–17, 2022, Philadelphia, PA, USA.* ACM, New York, NY, USA, 14 pages. https://doi.org/10.1145/3514221.3526125

SIGMOD '22, June 12–17, 2022, Philadelphia, PA, USA

ACM ISBN 978-1-4503-9249-5/22/06…$15.00
https://doi.org/10.1145/3514221.3526125

1 Introduction

Database-backed web applications have been the backbone of Internet applications from online shopping to banking. For many web applications, the database query latency is critical for the user experience. For example, it has been reported that an increase of 500ms in latency can reduce the traffic of a website by 20% [28], and users often give up a website when the loading time takes more than three seconds [41].

Query rewriting, which transforms an original query into a semantically equivalent alternative query [16–18], is an important step in query optimization. Effective rewrites can accelerate the execution time of input queries by orders of magnitude [6]. Rewriting relies on rules that specify the equivalence relations between queries. Existing rules are typically crafted by human experts and can take decades to accumulate [16–18, 27, 36, 37, 42, 44].

However, it is insufficient to rely on manual efforts to discover rewrite rules. The rich feature and subtle semantics of the query language make it challenging to prove equivalence [7, 10, 20] and to craft rules. As a result, the set of hand-written rewrite rules grows very slowly and misses many rewrite opportunities. The situation is made worse by the prevalent use of object-relational-mapping (ORM) frameworks in web application development. ORM frees programmers from explicitly constructing SQL queries but also results in non-intuitive queries whose patterns evade rules crafted by humans. To understand the impact of missed rewrites, we studied 50 real-world queries in several popular open-source web applications on Github. All of these queries have been rewritten by the developers to fix their performance issues. Even the latest version of SQL Server fails to rewrite 27 of these queries (54%) to a more efficient form as fixed manually by the developers. One such query incurs latency up to 37 seconds, while its equivalent rewritten one only takes 0.3 seconds [21] (details in §2.2).

In this paper, we propose WeTune, a rule generator that can automatically discover new rewrite rules without any human effort. Drawing inspiration from compiler superoptimization [2, 33], which finds a semantically equivalent optimal code sequence through the exhaustive search, WeTune aims to discover rewrite rules automatically via brute-force enumeration of all potential rules followed by a correctness check of each generated rule. During this discovery process, WeTune relies on heuristics to filter out those rules that are unlikely to improve performance, aka rules whose rewritten query contains more operators of each type than the original query. The remaining rewrite rules are deemed promising. We empirically determine the usefulness of these promising rules by using them to rewrite real world queries and measure the performance benefits of rewritten queries over synthetically generated database tables.

Those that result in beneficial rewrites are useful rules discovered by WeTune.

Although the high-level approach is simple, there are several challenges. First, how to represent rewrite rules in a general form that allows enumeration? Second, how to automatically verify the correctness of enumerated rules without human effort? To address these challenges, WeTune represents a rewrite rule as a pair of query plan templates together with a set of constraints that relate the templates to each other. It enumerates all possible query plan templates up to a threshold number of operators. A query plan template is generic in that it uses symbols instead of concrete names to represent tables, columns and predicates. WeTune further enumerates all constraints, which are conditions that could potentially make a pair of enumerated plan templates semantically equivalent. For instance, specific input relations of the two queries could be constrained to be the same, or the attributes used in a projection are restricted to be a subset of attributes in a certain relation. WeTune verifies the correctness of each rewrite rule using the SQL verifiers. It includes a built-in verifier, which provides a formal way of modeling rewrite rules as SMT formulas. Then the correctness problem can be automatically solved with an SMT solver. Besides the built-in verifier, WeTune also can support using existing SQL verifiers such as SPES to prove the correctness.

We have evaluated the effectiveness of WeTune on real-world database-backed applications using the 20 most popular web applications on GitHub. WeTune outputs 1106 promising rewrite rules, 35 of which are used to optimize queries of these applications. Furthermore, our results show that WeTune can successfully optimize 247 queries that are missed by existing systems, resulting in a latency reduction of up to 99%. Such optimization is due to WeTune's ability to discover new rewrite rules not known to any of the existing systems. WeTune can successfully verify the discovered rewrite rules.

In summary, our work makes the following contributions:

- A study showing that existing manually-discovered rewrite rules are insufficient for real-world queries in popular web applications.
- A demonstration that the enumeration approach introduced in compiler superoptimization can be applied in databases to generate query rewrite rules automatically.
- The formal modeling of database query rewrite rules to allow encoding to SMT formulas, which allows WeTune to verify the correctness of new rewrite rules.
- An evaluation of WeTune on a variety of real-world applications, which shows that it can successfully optimize 247 queries with substantial performance improvement.

2 Motivation

In this section, we examine the rewriting opportunities that existing commercial and open-source databases miss. Then, we study the impact of missed rewrites which cause developers to change a query into a more efficient form manually.

2.1 Insufficiency of Existing Rewrites

Existing databases already use a large number of rules to perform rewrites [16–18, 27, 36, 37, 42, 44]. Nevertheless, we have found that many queries still fail to be rewritten into a more efficient form by existing rules. This finding might come as a surprise: after all, decades of efforts have been spent on crafting rewrite rules, should not most—if not all—rules have been discovered already?

Original Query	Opt. By Existing DB	Ideal (WeTune)
q0: SELECT * FROM labels WHERE id IN (SELECT id FROM labels WHERE id IN (SELECT id FROM labels WHERE project_id=10) ORDER BY title ASC)	q1: SELECT * FROM labels WHERE id IN (SELECT id FROM labels WHERE project_id=10)	q2: SELECT * FROM labels WHERE project_id=10
q3: SELECT id FROM notes WHERE type='D' AND id IN (SELECT id FROM notes WHERE commit_id=7)	Unchanged	q4: SELECT id FROM notes WHERE type='D' AND commit_id=7

Table 1: Examples of the counter-intuitive queries generated by the ORM framework found in GitLab. The first column lists the original queries. The second one lists the best optimization results of the existing DB systems. The third one lists the ideal results, which can be achieved with the rules generated by WeTune. labels.id and notes.id are the primary keys of the tables, respectively.

To see why manual rewrite rules are insufficient, we want to keep in mind that these rules, more or less, are designed for SQL queries written directly by programmers. Human-written queries usually follow intuitive patterns which can be manually analyzed to distill useful rules. However, in modern web development, programmers no longer explicitly construct SQL queries. Rather, they typically make use of an *object-relational-mapping* (ORM) framework, which allows them to write object-oriented code to manipulate contents in the database. The underlying framework automatically generates SQL queries based on application logic. Not only are the resulting SQL queries generated by ORM opaque to programmers, but also they can be *counter-intuitive* to human rule developers. Table 1 shows two examples from the web application GitLab [15], a popular open-source version management website. Both SQL queries were generated by the ORM framework (ActiveRecord) as the result of running developers' Ruby code.

The first query $q0$ aims to select all the git merge requests whose project_id is "10". Specifically, it uses a subquery to compute the set of id values whose project_id is "10" according to the labels table. It then selects all rows from the labels table whose id falls within this set of subquery values. This query is counter-intuitive and inefficient in two places. First, the subquery to compute matching ids contains another inner subquery and the two subqueries are almost *identical*. Second, the ORDER BY clause of the inner subquery is unnecessary because the outermost IN operator treats the subquery SELECT...ORDER BY as an unordered list [1]. Ideally, the redundancies in the query should be identified by the query optimizer via a rewrite rule, so that the resulting optimized query resembles $q2$ as shown in Table 1. However, among MySQL, PostgreSQL, and MS SQL Server, only PostgreSQL and MS SQL Server can partially rewrite the query to $q1$ which removes ORDER BY and one of the two subqueries.

[1] Such behavior follows the SQL standard [22], and is confirmed in MySQL, PostgreSQL and Oracle DB. MS SQL Server explicitly denies such a query and reports the error "ORDER BY is disallowed in subquery".

Another query in Table 1 (q3) fetches id values from the notes table whose type is "D" and column_id is "7". For this query, the IN-selection is redundant because (1) the table used in the subquery is identical to the table used outside, which is the notes table; (2) The column projected by the subquery is the same as the column used in IN-selection, which is the primary key of the notes table. Hence, the subquery can be eliminated and the query could be transformed into a simple query as $q4$. Unfortunately, all three existing databases (MySQL, PostgreSQL and MS SQL Server) miss such opportunities and keep the query unchanged.

The above inefficient structures are unlikely seen in human-generated queries but are common for queries generated by ORM. Since the ORM-generated query results from running application code in different program locations or even third-party libraries, developers are agnostic to potential redundancies (e.g., duplicate subqueries) and inefficiency (e.g., unnecessary ORDER BY). Therefore, it is very difficult for developers to identify and fix the resulting performance issues.

2.2 Impact and Scope of Missed Rewrites

To better understand the impact of missed rewrites on real-world queries, we studied GitHub issues related to query performance in several popular web applications, including Discourse (discussion forum), GitLab (code management) and Spree (e-commerce) etc. Some are written in Ruby, while others are in Java (the full list can be found in our extended version [49]). The applications are chosen based on popularity, judged by the number of stars on GitHub [11, 13, 15, 25, 39, 40, 43, 46].

We manually inspected 50 GitHub issues related to query performance, with 15 from Discourse [25], 25 from GitLab [15], 4 from Spree [11], 2 from Redmine [43], and 4 from others [13, 39, 40, 46]. For all the 50 queries in our study, developers have fixed them by manually rewriting the original SQL query into a more efficient form, as the databases used by the application failed to rewrite these queries in the same efficient way. We have investigated whether state-of-the-art optimizers in different databases can rewrite these queries. Among these 50 queries, 27 queries (54%) cannot be rewritten into the desired forms in issues by the *latest* version of SQL Server (7 of them are similar to the examples in Table 1). The rewriters in MySQL, PostgreSQL, and Apache Calcite (including both Hep and Volcano Planner) perform even worse, failing to rewrite 38 (76%), 41 (82%), 47 (94%) and 46 (92%) of these queries, respectively. Our study shows that, although opportunities exist for many existing queries to be rewritten to a more efficient form, state-of-the-art manually curated rules miss such rewrites.

Among the Github issues in our study, a few of them [4, 21, 35, 48] give concrete numbers on the performance impact after the manual rewrite. For the example in [21], the original query latency can be up to 37 seconds, while the manually rewritten query only takes 0.3 seconds. Such latency difference is due to the optimizer failing to replace an IN-subquery with an INNER JOIN, which prevents the optimizer from selecting a better access path. The other issues [4, 35, 48] also lead to the 75%-99% latency reduction for their respective applications. Unfortunately, it is difficult to diagnose and resolve these performance problems. In particular, for these 50 issues, it took 13 months on average to fix one (via manual query rewrite). As developers do not directly write SQL but access the database via an ORM, they have less visibility and control over the final queries.

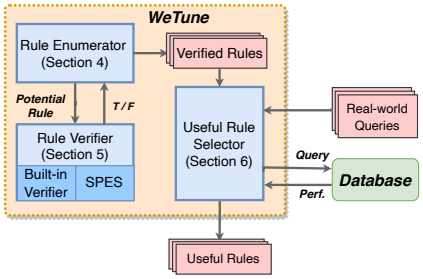


Figure 1: The architecture of WeTune.

3 Our Approach

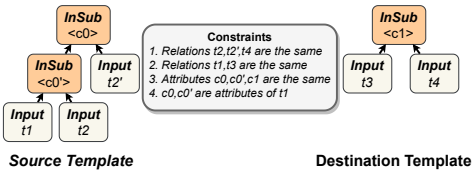
Manually crafted rewrite rules are no longer sufficient in an era where queries are automatically generated by web frameworks. To optimize these auto-generated queries, we need an automated approach to discover rewrite rules.

Basic Idea. WeTune aims to automatically discover new useful rewrite rules without any human effort. It is inspired by compiler superoptimization, especially peephole optimizer [2]. The peephole optimizer aims to transform a sequence of machine instructions into another equivalent but faster sequence and thus has a similar high-level goal as a database query optimizer. Peephole optimizers can automatically discover optimization rules via some form of brute-force search for the instruction sequences [2]. Inspired by this approach, we propose to automatically discover promising query rewrite rules through simple brute-force enumeration and to ensure the correctness of discovered rules through verification.

More concretely, WeTune's search for useful rewrite rules proceeds in two stages. In the first stage, WeTune discovers promising rules by enumerating the potential rewrite rules with the **Rule Enumerator** (§ 4) and verifying their correctness using the **Rule Verifier** (§ 5). We propose a new SMT-based verifier but WeTune can also use other verifiers (e.g. SPES [50]). In this stage, WeTune uses simple heuristics to filter out those rules that are unlikely to bring performance improvement; only promising rules are kept. In the second stage, WeTune empirically determines the usefulness of promising rules by using them to rewrite real-world queries and measuring the performance of the rewritten queries (§ 6). Figure 1 shows the overall architecture of WeTune.

Our high-level approach is straightforward. However, to make it work, we must resolve several technical challenges that face rule enumeration and verification. These challenges are unique to query rewriting and not present in compiler optimization.

(1) How to represent a rewrite rule to make it amicable to enumeration? A rule consists of a pair of queries, which must be generic and not bound to concrete tables and columns. How to enumerate generic queries and make a source query equivalent to a destination query? (Section 4)
(2) How to determine whether an enumerated rewrite rule is correct? Can we adopt an existing query equivalence checker that requires concrete queries to work with generic queries? Can




```
q5: ... FROM T WHERE T.x IN (SELECT R.y FROM R)
                   AND T.x IN (SELECT R.y FROM R)
q6: ... FROM T WHERE T.x IN (SELECT R.y FROM R)
```

Figure 2: An example rule found by WeTune (No.4 in Table 7). It can eliminate redundant IN-subquery operator of a SQL query such as $q5$, and rewrite it into $q6$. Existing databases miss the opportunity to rewrite such a counter-intuitive queries.

we develop a new verifier to address the limitations of existing checkers? (Section 5)

4 Rule Enumerator

WeTune models the rewrite rule as a triple $<q_{src}, q_{dest}, C>$, where q_{src} is a source query plan template, q_{dest} is a destination query plan template and C is a set of constraints. A query plan template is a fragment of the logical query plan tree whose operators include selection, projection, etc. Unlike those in a concrete query, the table names, attributes and predicates in a query plan template are symbolic. The constraint set C consists of a set of predicates, each of which describes some relationship between the symbols from the source and destination query plan templates. The rule specifies that if all constraints in C are satisfied, then q_{src} and q_{dest} are semantically equivalent. Given a SQL query q, if some fragment in q matches q_{src}, the matched fragment can be replaced with the corresponding fragment q_{dest} that satisfies C.

Figure 2 shows an example rewrite rule which can eliminate redundant IN-subquery in a SQL query. The source template q_{src} is represented as $\mathsf{InSub}_{c0}(\mathsf{InSub}_{c0'}(\mathsf{t1}, \mathsf{t2}), \mathsf{t2'})$. The operator InSub_{c0} has a left child $\mathsf{InSub}_{c0'}$ and a right child t2'. InSub is an operator in the query plan template which represents IN-subquery. It represents the queries (e.g., $q5$) with two IN-subquery operators, and these two IN-subquery are connected by AND. The destination template q_{dest} is $\mathsf{InSub}_{c1}(\mathsf{t3}, \mathsf{t4})$. The constraint set C specifies the following constraints: $t2$, $t2'$ and $t4$ are the same relations; $t1$ and $t3$ are the same relations; $c0$, $c0'$ and $c1$ are the same attributes. The figure also shows a SQL query $q5$ derived from Gitlab [15]. This query matches q_{src}. Thus, it can be replaced by a better query $q6$ which follows the pattern specified by q_{dest} under the constraints in C. This inefficiency pattern is quite counter-intuitive as its two inner subqueries are almost identical. However, none of our studied DBs can successfully optimize this query.

The Rule Enumerator enumerates potential rewrite rules. To do so, it first enumerates all possible plan templates (Section 4.1). To restrict the search space, it bounds the template size so that the number of operators in a template is within some small threshold. Then, for every pair of plan templates, it enumerates all potential constraints (Section 4.2). Last, it selects the promising rules which are likely to improve the query performance (Section 4.3).

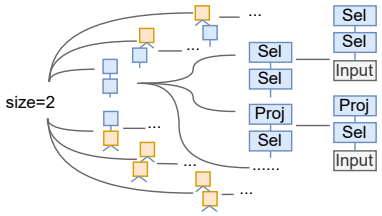


Figure 3: The example of enumerating query plan templates. A blue block denotes an operator with one input, while a yellow block denotes an operator with two inputs.

4.1 Plan Template Enumerator

The query plan template is a tree whose nodes are relational algebra operators with symbolic inputs or parameters [16, 18]:

Operator. Each operator takes one or two relations as input (except Input operator itself), performs algebraic computation according to its semantic, and outputs a single relation. Currently, WeTune only supports the operators in Table 2.

Symbol. In a concrete query plan, operators can be parameterized by concrete schema information such as column names, etc. In a query plan template, such concrete parameters are replaced with symbolic ones. There are three kinds of symbols:

- *Relation Symbol.* A relation symbol *rel* (r for short) represents a relation. It is used to parameterize the input relation of a plan template. e.g., t1, t2, t2', t3 and t4 in Figure 2.
- *Attribute list Symbol.* An attribute list symbol *attrs* (a for short) represents a sequence of attributes. In Figure 2, c0, c0' and c1 are attribute lists. Each of them contains at least one attribute. Additionally, each relation symbol r is associated with an attribute list symbol a_r that represents all the attributes in r.
- *Predicate Symbol.* A predicate symbol *pred* (p for short) represents a predicate, which takes zero or more values as input and yields a boolean value. It is used to parameterize the predicate expression of the selection operator.

The enumeration strategy of WeTune separately enumerates a query plan's tree structure and the operator types for each tree node. More concretely, the enumeration is done in three steps: first, WeTune constructs all possible tree structures with two kinds of internal tree nodes: one type is the node having one child, and another type is the node having two children; Second, for each tree structure, it exhaustively assigns the operators listed in Table 2 to every node to enumerate concrete trees. The number of operator's inputs should match the number of the node's children; Last, it adds Input nodes as the leaf nodes' children. Figure 3 shows the process of enumerating query plan templates having two operators. To reduce the enumeration space, WeTune only enumerates templates up to 4 operators excluding the Input nodes. Furthermore, it filters out those templates that lead to an invalid SQL query, such as misplacing the Deduplication operator.

4.2 Constraint Enumerator

WeTune pairs the enumerated templates as $<q_{src}, q_{dest}>$, and searches for the constraint set that would turn the pair of templates into a valid rewrite rule. A constraint is a predicate that specifies the relationship between symbols in q_{src} and q_{dest}. To bound the search

Operator Name	Symbol	#Input	Description
Input	Input_r	0	$\text{Input}_r()$ represents an initial input relation specified by r.
Projection	Proj_a	1	$\text{Proj}_a(R)$ projects its input relation R on attributes specified by a.
Selection	$\text{Sel}_{p,a}$	1	$\text{Sel}_{p,a}(R)$ discards tuples in its input relation R that do not satisfy the predicate p. Values on attributes a from R are used to evaluate the predicate p.
In-Sub Selection	InSub_a	2	$\text{InSub}_a(R_l, R_r)$ discards tuples in the left input R_l that are absent in the right input R_r. Values on attributes a from R_l are used for the presence check.
(Inner/Left/Right) Join	$(\text{I/L/R})\text{Join}_{a_l,a_r}$	2	$\text{IJoin}_{a_l,a_r}(R_l, R_r)$ Cartesian products its input relations R_l and R_r, then discards the tuples that have mismatched values on attributes a_l and a_r. (L/R)Join additionally keeps the mismatched tuples and fills NULL on the right/left-side attributes.
Deduplication	Dedup	1	Dedup(R) discards duplication of tuples in its input relation R.

Table 2: SQL operators supported by WeTune.

space, we consider the following limited set of constraints, drawn from our experience of studying existing rewrite rules and examining developers' manual query rewrites.

- *RelEq*(rel_1, rel_2). This constraint indicates that two relation symbols, rel_1 and rel_2, are equivalent (i.e., contain the same tuples).
- *AttrsEq*($attrs_1$, $attrs_2$). This constraint indicates that two attribute list symbols, $attrs_1$ and $attrs_2$, have the same sequence of attributes.
- *PredEq*($pred_1$, $pred_2$). This constraint indicates that two predicate symbols, $pred_1$ and $pred_2$, are equivalent (i.e. $pred_1 \Leftrightarrow pred_2$).
- *SubAttrs*($attrs_1$, $attrs_2$). This constraint indicates that each attribute in $attrs_1$ is also in $attrs_2$. It can be used to express which relations an attribute list is from. For example, in Figure 2, *SubAttrs*(c_0, a_{t1}) indicates that each attribute in c_0 corresponds to some column from table $t1$.
- *RefAttrs*(rel_1, $attrs_1$, rel_2, $attrs_2$). This constraint indicates that any value in the relation rel_1 on the attribute $attrs_1$ is also in rel_2 on $attrs_2$.
- *Unique*(*rel*, *attrs*). This constraint indicates that every value in *rel* on *attrs* is unique.
- *NotNull*(*rel*, *attrs*). This constraint indicates that every value in *rel* on *attrs* is not NULL.

Given a pair of plan templates $<q_{src}, q_{dest}>$, constraint enumeration generates the set C^*, which contains all possible constraints related to q_{src} and q_{dest}. This is done by exhaustively filling in the parameters of constraints above with the symbols in q_{src} and q_{dest}. We use C^* later in the search for promising rules.

4.3 Searching for Promising Rules

Given a pair of query templates $<q_{src}, q_{dest}>$, and constraint set C^* which includes all constraints related to q_{src} and q_{dest}, WeTune needs to search for some subset C of C^* which makes q_{src} and q_{dest} semantically equivalent. Furthermore, we keep only those valid rules which are deemed promising.

A rule of $<q_{src}, q_{dest}, C>$ is promising if it satisfies the following two requirements: first, C is the most relaxed constraint set, such that the removal of any constraint in C compromises the correctness of the rule. In other words, C is the minimal constraint set that enables the equivalence between q_{src} and q_{dest}; Second, q_{dest} does not have more operators of each type than q_{src}. With this heuristic, rewrite rules will simplify rather than complicate the source query, thus are more likely to improve the query performance.

Algorithm 1: Search for Promising Rules

```
1  EnumerateRules(k):
2    T := EnumerateTemplates(k)
3    R := ∅
4    foreach <q_src, q_dest> ∈ T × T do
5      if q_dest is not simpler than q_src then continue
6      C* := EnumerateConstraints(q_src, q_dest)
7      ℂ := SearchRelaxed(q_src, q_dest, C*)
8      R := R ∪ {<q_src, q_dest, C> | C ∈ ℂ}
9    return R
10 SearchRelaxed(q_src, q_dest, C*):
11   if ¬ProveEq(q_src, q_dest, C*) then return ∅
12   ℂ := ∅
13   foreach c ∈ C* do
14     ℂ := ℂ ∪ SearchRelaxed(q_src, q_dest, C* − {c})
15   if ℂ = ∅ then return {C*}
16   else return ℂ
```

Algorithm 1 shows the basic algorithm to search for promising rules. It first enumerates all query templates, as described in Section 4.1. Then, it pairs the enumerated templates as $<q_{src}, q_{dest}>$ and keeps those whose q_{dest} has the same or fewer operators of each type than q_{src} (Line 5). For each pair $<q_{src}, q_{dest}>$, it generates constraint set C^* by enumerating all possible constraints related to q_{src} and q_{dest}. Last, it invokes `SearchRelaxed` to recursively search for the subsets of C^* to form the promising rules.

The function `SearchRelaxed` starts with C^*, and iteratively relaxes the constraint set by removing one constraint (Line 14) and verifies the resulting rule correctness (Line 11). Specifically, it uses an underlying verifier to prove the equivalence between q_{src} and q_{dest} under the constraints in C (Section 5). If the verification fails, we know the constraint set is too relaxed to imply the equivalence. In this case, we stop further relaxing and traceback (Line 11). If no constraint can be further removed, we have found the most relaxed constraint set (s) (Line 15). Note that there could be multiple most relaxed sets, and `SearchRelaxed` tries to find all of them. This is why it returns a set of sets, and each member is a most relaxed set.

To reduce the search cost, WeTune introduces the following methods: first, it excludes the useless constraints from C^*. A constraint is considered useless if it only involves the symbols in q_{dest} or leads to an illegal query plan; Second, instead of examining every subset of C^*, it only checks the subsets which are both closures and non-conflicting. A subset is a closure if it cannot imply any constraint absent from the set. Meanwhile, a subset is non-conflicting

if no constraints in the subset conflict with each other. Two constraints have a conflict if putting them together will introduce an illegal plan. Last, WeTune skips checking the constraint set C if it can be implied by a constraint set C', and C' has already been proved to make q_{src} and q_{dest} equivalent.

5 Rule Verifier

WeTune depends on the rule verifier to ensure correctness. A rule $<q_{src}, q_{dest}, C>$ is correct if the source and destination query plan templates are semantically equivalent when the constraints hold. We design a rule verifier based on FOL (first-order logic) formulation (Section 5.1). WeTune can also use an existing SQL equivalence checker such as SPES [50] to verify rules (Section 5.2).

5.1 Built-in Rule Verifier

At the high level, WeTune's built-in verifier works by first representing a rule $<q_{src}, q_{dest}, C>$ as a U-semiring expression [8], and then converting the expression into FOL formulas. Finally, the FOL formulas are verified using an SMT solver.

5.1.1 *Formal Representation of Rules*

Given a rewrite rule, we use U-expression [8] to represent q_{src} and q_{dest}, and use FOL formulas to specify the constraint set C.

U-expression. Inspired by UDP [8], we also use U-expressions to model SQL queries under the bag semantics, which capture the multiplicity of a tuple in the relation. Under this representation, a query is viewed as operations on a semiring of natural numbers [8, 19]. We adopt the terms defined in UDP [8], which are summarized below:

- $[\![R]\!](x)$ returns the multiplicity of the tuple x in the relation R.
- $[b] \triangleq$ if b then 1 else 0. Since this expression converts a boolean value to an integer, it can be used to turn a predicate into a U-expression.
- $||e|| \triangleq$ if $e > 0$ then 1 else 0, where e is a U-expression. It models `Deduplication`.
- $not(e) \triangleq$ if $e > 0$ then 0 else 1, where e is a U-expression. It models the negation of a predicate.
- $\sum_{t \in D} f(t) \triangleq f(t_0) + f(t_1) + \cdots$, for all $t_i \in D$, where D is a tuple set called *summation domain*, and f is a function $D \to N$. By default, D is an infinite set containing all possible tuples. $\sum_{t \in D} f(t)$ models `Projection`.

In order to model a query plan made out of different operators, these terms are connected by "+" and "∗", which have the same meaning as that of natural numbers. For example, we can use $||\sum_x [t = x.k] \times [\![R]\!](x) \times [x.a > 12]||$ to denote the multiplicity of the tuple t in the output relation of "SELECT DISTINCT x.k FROM R AS x WHERE x.a>12". The summation can be omitted if the projection retains all attributes. For example, $||[\![R]\!](t) \times [t.a > 12]||$ can represent the multiplicity of the tuple t in the output relation of "SELECT DISTINCT * FROM R AS x WHERE x.a > 12". In the following paragraphs, we omit D in summation and just write $\sum_t f(t)$, where t is the *summed variable* of the summation.

Converting the query template into U-expression. WeTune translates each query template q to a function $[\![q]\!](t) : Tuple \to \mathbb{N}$, which takes a tuple t as input and returns its multiplicity in the output relation of the query template. The multiplicity is represented as a U-expression. Unlike UDP [8] which performs the translation to U-expression for concrete queries, WeTune translates for symbolic query templates. The translation involves two steps:

Operator	Expression
Input_r	$f(t) \coloneqq r(t)$
Proj_a	$f(t) \coloneqq \sum_{t_l}(f_l(t_l) \times [t = a(t_l)])$
$\text{Sel}_{p,a}$	$f(t) \coloneqq f_l(t) \times [p(a(t))]$
InSubSel_a	$f(t) \coloneqq f_l(t) \times \|\|f_r(a(t))\|\| \times not([IsNull(a(t))])$
IJoin_{a_l,a_r}	$f(t) \coloneqq \sum_{t_l,t_r}([t = t_l \cdot t_r] \times f_l(t_l) \times f_r(t_r) \times [a_l(t_l) = a_r(t_r)] \times not([IsNull(a_l(t_l))]))$
LJoin_{a_l,a_r}	$f(t) \coloneqq$ (IJoin Expr.) $+ \sum_{t_l,t_r}([t = t_l \cdot t_r] \times f_l(t_l) \times [IsNull(t_r)] \times not(\sum_{t'_r}(f_r(t'_r) \times [a_l(t_l) = a_r(t'_r)] \times not([IsNull(a_l(t_l))]))))$
RJoin_{a_l,a_r}	(symmetric to LJoin)
Dedup	$f(t) \coloneqq \|\|f_l(t)\|\|$

Table 3: The rules to translate the SQL operator into U-expression. Each U-expression is a function taking a tuple t and returning its multiplicity in the relation produced by the operator. f_l and f_r represent the U-expressions of the operator's left and right children, respectively. $t = t_l \cdot t_r$ is a predicate requiring t is the concatenation of t_l and t_r.

Example SQL
$q5$: `... FROM T WHERE T.x IN (S) AND T.c IN (S)`
$q6$: `... FROM T WHERE T.x IN (S)`

Templates
q_{src}: $InSub_a(InSub_a(r_0,\ r_1),\ r_1)$
q_{dest}: $InSub_a(r_0,\ r_1)$

$$[\![q_{src}]\!](t) \coloneqq r_0(t) \times not([IsNull(a(t))]) \times ||\sum_x r_1(x) \times [x = a(t)]|| \times not([IsNull(a(t))]) \times ||\sum_x r_1(x) \times [x = a(t)]||$$

$$[\![q_{dest}]\!](t) \coloneqq r_0(t) \times not([IsNull(a(t))]) \times ||\sum_x r_1(x) \times [x = a(t)]||$$

Figure 4: The U-expressions of the rewrite rule in Figure 2. S denotes the entire subquery of "SELECT R.y FROM R". **q_{src} and q_{dest} are the source and destination templates. The symbols of a, r_0 and r_1 in the templates represent** T.x, T, **and the relation produced by** S **in the example SQL accordingly.**

Step 1. Translating the symbols in the query template. We use *uninterpreted* functions to enable the translation:

- Each relation symbol *rel* corresponds to a function $[\![r]\!](t) : Tuple \to \mathbb{N}$ that takes a tuple t as input and returns its multiplicity in *rel*.
- Each attribute list symbol *attrs* corresponds to a function $[\![a]\!](t) : Tuple \to Tuple$ that takes a tuple t as input, projects it on the attributes in *attrs* and returns the projected tuple.
- Each predicate symbol *pred* corresponds to a function $[\![p]\!](t) : Tuple \to Bool$ that takes a tuple t as input and returns whether the tuple satisfies *pred*.

For brevity, we omit $[\![\ ]\!]$ whenever there is no ambiguity. For example, $r(t)$ denotes the application of a relation function.

Step 2. translating the plan structure. This is done by recursion on the tree structure, as depicted by Algorithm 2. The function `ToUExpr` takes a (sub-)plan template as input. It returns the translated expression and a representative tuple of the output relation. For each operator, the algorithm recursively calculates the expressions of its children, then looks up in Table 3 to build its own expression based on its children's expressions. Figure 4 shows the

translated U-expression of the example in Figure 2. We will discuss the *IsNull* predicate next.

Algorithm 2: Translate Plan to U-expression

```
1 ToUExpr(q):
2   <f_l, t_l> := ToUExpr(q.child[0]) //None if no child
3   <f_r, t_r> := ToUExpr(q.child[1]) //None if single child
4   return TranslateByTable3(q, f_l, t_l, f_r, t_r)
```

Handling NULL. One of the biggest limitation of UDP's modeling of SQL query is its assumption that none of the attributes in a relation is NULL. Consequently, UDP cannot support the OUTER JOIN operator. According to our study, more than half of SQL queries collected from the web application involve such operators. To handle both NULL and OUTER JOIN, WeTune's translation of U-expression takes into consideration the impact of NULL on the operators, as shown in Table 3. Supporting other operators such as aggregation is trickier and left as future work.

For operator Input_r, the expression r(NULL) returns the multiplicity of NULL tuples [2] in the input relation; For Proj_a, f(NULL) = $\sum_{t_l}(f_l(t_l) \times [\text{NULL} = a(t_l)])$ will return the multiplicity of tuples from t_l whose attribute a is NULL; For $\text{Sel}_{p,a}$, some predicate p such as ">" and "<" will return unknown on evaluating NULL. When $a(t)$ is NULL and $p(a(t))$ returns *unknown*, $[p(a(t))]$ will return 0, which is the same as $[false]$. Here, we are able to treat the *unknown* in the three-valued logic as the *false* in two-valued logic, as $\text{Sel}_{p,a}$ only evaluates the tuple that makes the predicate $true$; For Dedup, it returns 1 if there is at least one NULL tuple in the relation, otherwise 0.

To model the impact of NULL on the other operators in Table 3, WeTune introduces a new predicate *IsNull* to U-expression. When x is NULL, $IsNull(x)$ returns true and $[IsNull(x)]$ is 1. With such predicate, WeTune is able to filter out the NULL tuples. In detail, for InSubSel_a (IN-subquery), it uses the *IsNull* predicate to filter out the NULL tuple from the outer query. For INNER JOIN, it uses the *IsNull* to filter out the cases that left or right relation has NULL tuples. We will discuss how to handle NULL for OUTER JOIN in the next paragraph.

Supporting Outer Join operators. WeTune supports the OUTER JOIN operator by using the specific rule in Table 3 based on the modeling of NULL. Unlike INNER JOIN, OUTER JOIN keeps the rows that do not have a matching row on the other side and fills the void with NULL. For example, "x LEFT JOIN y ON x.a = y.b" keeps all rows from the left table x. For a left row that does not match any right row on x.a = y.b, NULL is appended as the right row. Hence, as shown in Table 3, the LEFT JOIN is the addition of two parts: (1) for those matched rows, the same as INNER JOIN; (2) for those non-matched rows, a product of three terms: "$f_l(t_l)$" describes the left rows being kept; "$[IsNull(t_r)]$" describes that NULL is appended as the right row; "$not(\sum_{t'_r}(\cdots))$" describes the non-matching condition. Figure 5 shows an example of translating a LEFT JOIN.

Representing constraints with FOL formulas. Each constraint is directly translated to a FOL formula according to Table 4.

[2] A tuple is NULL if all its attributes are NULL. A NULL attribute can be considered as a NULL tuple with only one attribute.

Constraint	Expression
$RelEq(r_1, r_2)$	$\forall t.r_1(t) = r_2(t)$
$AttrsEq(a_1, a_2)$	$\forall t.a_1(t) = a_2(t)$
$PredEq(p_1, p_2)$	$\forall t.p_1(t) = p_2(t)$
$SubAttrs(a_1, a_2)$	$\forall t.a_1(t) = a_1(a_2(t))$
$RefAttrs(r_1, a_1, r_2, a_2)$	$\forall t_1.((r_1(t_1) > 0 \wedge \neg(IsNull(a_1(t_1)))) \Rightarrow \exists t_2.(r_2(t_2) > 0 \wedge \neg(IsNull(a_2(t_2))) \wedge [a_1(t_1) = a_2(t_2)]))$
$Unique(r, a)$	$(\forall t.r(t) \leq 1) \wedge (\forall t, t'.r(t) > 0 \wedge r(t') > 0 \wedge a(t) = a(t') \Rightarrow t = t')$
$NotNull(r, a)$	$\forall t.r(t) > 0 \Rightarrow \neg(IsNull(a(t)))$

Table 4: Translation table from constraint to FOL formulas.

A set of constraints C is translated to the conjunction of its members:

$$\text{ToFOL}(C) \triangleq \bigwedge_{c \in C} \text{ToFOL}(c)$$

5.1.2 *Verification of the Rule Correctness*

After formalizing the query templates with U-expressions and the constraints with FOL formulas, the rule verifier will check a rule's correctness using the SMT solver. To do so, we need to formalize the correctness of the rule with FOL formulas.

Defining a rule's correctness. To formalize correctness, we need to first introduce the concept of *interpretation*, which specifies the meaning of the relation, predicate and attribute list symbols.

Definition 1 (Interpretation)**.** Given a query plan template q represented as a U-expression, an interpretation is an assignment of meaning to all symbols in q. We denote the concrete query plans under the interpretation I by q^I. Similarly, the truth value of a constraint set C under I is denoted by C^I.

Next, we define the correctness of a rule. Intuitively, if a rule is correct, its source and destination query templates should be equivalent under the rule's constraint set for any interpretations.

Definition 2 (Correctness of a rewrite rule)**.** Given a rule with two query plan templates $<q_{src}, q_{dest}>$ and a constraint set C, q_{src} and q_{dest} are equivalent under C iff the following formula holds.

$$\forall I.C^I \Rightarrow \forall t.q^I_{src}(t) = q^I_{dest}(t)$$

The formula "$\forall t.q^I_{src}(t) = q^I_{dest}(t)$" is consistent with the definition of bag equivalence [19]: two bags are equivalent iff. every tuple has the same multiplicity on both sides. Moreover, the outer quantifier "$\forall I$" requires the proposition to hold under any interpretation.

To prove query equivalence, UDP [8] relies on converting two U-expressions to their normalized forms and then establishing syntactic isomorphism between them. However, such syntactic isomorphism requires establishing a one-to-one equivalent relationship between the summations in the U-expressions, which can not be guaranteed for queries with operators like OUTER JOIN. Figure 5 shows an example. Since the two normalized expressions $[\![q_{src}]\!](t)$ and $[\![q_{dest}]\!](t)$ have different numbers of summations, UDP cannot establish the isomorphism needed for proving equivalence.

Logic-based decision procedure. Unlike UDP, WeTune uses a logic-based decision procedure, which translates the correctness definition (Definition 2) to a FOL formula and verifies it with the SMT solver. There are two challenges in realizing this approach.

Example SQL $q7$: `SELECT T.* FROM T LEFT JOIN S ON T.k=S.k'`
$q8$: `SELECT T.* FROM T`
`Integrity Constraint: S.k' is unique key`

Templates q_{src}: $Proj(LJoin_{a_0,a_1}(r_0,\ r_1))$
q_{dest}: $Proj(r_0)$

$$[\![q_{src}]\!](t) \coloneqq \sum_{x,y}([t = x] \times r_0(x) \times r_1(y) \times [a_0(x) = a_1(y)] \times NonNull(a_0(x))) + \sum_{x,y}([t = x] \times r_0(x) \times [IsNull(y)] \times not(\sum_{y'} r_1(y') \times [a_0(x) = a_1(y')] \times NonNull(a_0(x))))$$

$$[\![q_{dest}]\!](t) \coloneqq \sum_{x}([t = x] \times r_0(x))$$

Figure 5: A pair of equivalent queries that cannot be proved by UDP. The SQL query of $q7$ is collected from an open sourced web application Discourse [25]. q_{src} and q_{dest} are the templates. The symbols of r_0, r_1, a_0 and a_1 in the template can represent the relations of T and S, and the attributes of T.k and S.k' in $q7$ accordingly. $NonNull(\cdot)$ is an abbreviation of $not([IsNull(\cdot)])$ for simplicity.

The first challenge is how to translate the U-expression $q_{src}(t) = q_{dest}(t)$ to a FOL formula. WeTune performs the translation according to Table 5. The table shows the basic U-expressions used by the translation and their corresponding FOL formulas. These FOL formulas ensure the sufficient condition, i.e. for any interpretation that satisfies the FOL formula, then it also satisfies the U-expression. For a compound U-expression, WeTune performs recursive translation. Starting from $q_{src}(t) = q_{dest}(t)$, which defines the correctness of a rewrite rule, WeTune individually translates $q_{src}(t)$ and $q_{dest}(t)$ into FOL formulas. When performing the translation, it will find the matched form in Table 5 and replace it with the FOL formula[3] For example, Figure 6 shows the translated FOL formula when proving the equivalence of two queries in Figure 2.

$$[\![q_{src}]\!](t) = r_0(t) \times [\neg(IsNull(a(t)))] \times [\exists x.r_1(x) \times [x = a(t)] > 0] \times [\exists x.r_1(x) \times [x = a(t)] > 0]$$

$$[\![q_{dest}]\!](t) = r_0(t) \times [\neg(IsNull(a(t)))] \times [\exists x.r_1(x) \times [x = a(t)] > 0]$$

Figure 6: The first-order logic formula of the example in Figure 2. $[\cdot]$ denotes the transformation from bool to natural number: $[b] \coloneqq ite(b, 1, 0)$.

When encoding the FOL formulas for the SMT solver, we represent the tuple as an object with uninterpreted sort in SMTLIB; The relation, represented as $[\![R]\!](t)$ in U-expression, is encoded as an uninterpreted function $R(t) : Tuple \rightarrow \mathbb{N}$; the predicate is encoded as an uninterpreted function $P(t) : Tuple \rightarrow Bool$.

When translating U-expressions to FOL formulas, the most difficult part is the translation of summation (the last two rows in Table 5). The unbounded summation domain makes it difficult to represent the value of a summation in a FOL formula. We address the problem based on the following insight. Since what matters is the equivalence relation, it is unnecessary to explicitly represent the value of a summation. Therefore, when proving $\sum_t f(t) = \sum_t f'(t)$, we aim to find the sufficient condition P such that $P \Rightarrow \sum_t f(t) = \sum_t f'(t)$. When P is proved to be true, then $\sum_t f(t) = \sum_t f'(t)$ must also hold. Furthermore, if such P does not involve summation, we can instead translate P into a FOL formula and prove it, effectively eliminating the summation.

When P is not true, the verification fails and we consider the rewrite rule to be incorrect, which can prevent an incorrect rule from passing the verification. Specifically, we propose Theorem 5.1 and Theorem 5.2, corresponding to the last two rows in Table 5.

Theorem 5.1 eliminates the summation when the summed variables of two summations are aligned. It can be generalized to multiple summed variables. The proof can be found in our extended version [49].

Theorem 5.1.

$$\left(\forall I \forall t. r^I(t) \times f^I(t) = r^I(t) \times g^I(t)\right) \Leftrightarrow \left(\forall I. \sum_t \left(r^I(t) \times f^I(t)\right) = \sum_t \left(r^I(t) \times g^I(t)\right)\right)$$

where r is a function that denotes a relation, $f(t)$ and $g(t)$ are arbitrary expressions. The superscript I indicates the interpretation of symbols under I.

Theorem 5.2 generalizes Theorem 5.1 to scenarios where the summed variables are not aligned. The proof can be found in our extended version [49].

Theorem 5.2.

$$\left(\forall I \forall t. \left(r^I(t) \times f^I(t) \neq r^I(t) \times g^I(t) \wedge r^I(t) \times f^I(t) = 0 \wedge \sum_s h^I(t,s) = 0\right) \vee \left(r^I(t) \times f^I(t) = r^I(t) \times g^I(t) \wedge \left(r^I(t) \times f^I(t) = 0 \vee \sum_s h^I(t,s) = 1\right)\right)\right) \Rightarrow \left(\forall I. \sum_t \left(r^I(t) \times f^I(t)\right) = \sum_{t,s} \left(r^I(t) \times g^I(t) \times h^I(t,s)\right)\right)$$

where $h(t,s)$ is an arbitrary expression.

The second challenge is that universal quantifiers may make the proof undecidable and cause the SMT solver to timeout. When

U-expression	FOL formula
$f_1(t) = f_2(t)$	$\mathrm{Tr}(f_1(t)) = \mathrm{Tr}(f_2(t))$
$f_1(t) + f_2(t)$	$\mathrm{Tr}(f_1(t)) + \mathrm{Tr}(f_2(t))$
$f_1(t) \times f_2(t)$	$\mathrm{Tr}(f_1(t)) \times \mathrm{Tr}(f_2(t))$
$\|\|f(t)\|\|$	$ite(\mathrm{Tr}(f(t)) > 0, 1, 0)$
$not(f(t))$	$ite(\mathrm{Tr}(f(t)) > 0, 0, 1)$
$[p]$	$ite(\mathrm{Tr}(p), 1, 0)$
$\|\|\sum_t f(t)\|\|$	$ite(\exists t.\mathrm{Tr}(f(t)) > 0, 1, 0)$
$not(\sum_t f(t))$	$ite(\exists t.\mathrm{Tr}(f(t)) > 0, 0, 1)$
$\sum_t f(t) = 0$	$\forall t.f(t) = 0$
$\sum_t f(t) = 1$	$\exists t.(f(t) = 1 \wedge (\forall t'.t' \neq t \Rightarrow f(t') = 0))$
$\sum_t r(t) \times f(t) = \sum_t r(t) \times g(t)$	$\forall t.r(t) \times \mathrm{Tr}(f(t)) = r(t) \times \mathrm{Tr}(g(t))$
$\sum_t r(t) \times f(t) = \sum_{t,s} r(t) \times g(t) \times h(t,s)$	$\forall t.\ ((r(t) \times \mathrm{Tr}(f(t)) \neq r(t) \times \mathrm{Tr}(g(t)) \wedge r(t) \times \mathrm{Tr}(f(t)) = 0 \wedge \mathrm{Tr}(\sum_s h(t,s) = 0)) \vee (r(t) \times \mathrm{Tr}(f(t)) = r(t) \times \mathrm{Tr}(g(t)) \wedge (r(t) \times \mathrm{Tr}(f(t)) = 0 \vee \mathrm{Tr}(\sum_s h(t,s) = 1))))$

Table 5: Translation table from U-expression to FOL formulas. Function Tr recursively translates sub-expressions according to this table. The $ite(p, 0, 1)$ means if p is true, then the formula returns 0. Otherwise, the formula returns 1.

[3] Occasionally, WeTune can not find any match in Table 5. For example, if both the left and right child of a left join operator are an IN-subquery, then the U-expression contains two sums (aka $\sum$), which cannot be converted into FOL by Table 5. In this case, the verifier cannot prove the rule's correctness.

proving a FOL formula is a tautology, the SMT solver needs to exhaustively check all cases. For example, to prove $q_{src}(t)$ is always equivalent to $q_{dest}(t)$, it needs to check all possible interpretations, and under each interpretation it needs to further check every tuple t. In contrast, it is much easier proving a FOL formula is unsatisfiable (UNSAT), as the SMT solver will stop as soon as it finds a contradiction implying UNSAT, which can avoid exhaustive reasoning.

Therefore, given a rewrite rule of $<q_{src}, q_{dest}, C>$, WeTune verifies its correctness by proving that $\neg(C \Rightarrow \forall t.q_{src}(t) = q_{dest}(t))$ is UNSAT. As a result, instead of exhaustively checking all possible interpretations and tuples, the SMT solver only needs to find a contradiction that implies the formula above is UNSAT to prove rule correctness. Nevertheless, timeouts still occur because the SMT solver may fail to find the contradiction when the rule is either incorrect or too complicated. To evaluate the effect of the timeout, we test 232 rewrite rules from Calcite test suite which are already known to be correct. WeTune can successfully prove 73 rules without timeout. The others cannot be proved because they involve the operators or features that WeTune does not support. We also generate 100 incorrect rules by randomly selecting rules of Calcite and mutating their constraints to make them incorrect. WeTune encounters timeout for 96 of them, and only 4 rules are successfully proved to be incorrect.

In summary, by converting the correctness reasoning to be the UNSAT problem, WeTune is likely to perform the reasoning without timeout when the rules are correct. Our empirical evidence suggests that, for incorrect rules, WeTune tends to encounter timeout instead of giving a counterexample. Thus, WeTune conservatively considers those rules which cause timeout to be incorrect. Currently, we only focus on finding the correct rules and leave checking the incorrect ones without timeout as future work.

5.2 Integrating SPES

WeTune can also use an existing query equivalence checker like SPES [50] to further improve rule discovery in scenarios when its built-in verifier in Section 5 cannot prove a rule's correctness.

Compared to the built-in verifier, SPES additionally supports UNION and Aggregation operators. Therefore, we extend the rule enumerator in Section 4.1 to enumerate plan templates containing these two operators. The Aggregation operator is parameterized with 4 symbols: an attribute list symbol a_{group} for attributes used in the GROUP BY clause; another attribute list symbol a_{agg} for attributes used in the aggregate function; an uninterpreted function symbol f for the aggregate function; a predicate symbol p for the predicate in HAVING clause. For example, the SQL query "SELECT a_{group}, $f(a_{agg})$ FROM ... GROUP BY a_{group} HAVING $p(a_{group})$" is represented as a plan template $Agg_{a_{group},a_{agg},f,p}(...)$. WeTune also adds a new constraint $AggrEq(f_1, f_2)$ to indicate that two aggregate functions are equivalent. For the UNION operator, WeTune does not introduce any new symbols or constraints.

Given a rewrite rule $<q_{src}, q_{dest}, C>$, WeTune needs to convert it into inputs accepted by SPES. As SPES only takes the concrete SQL queries and does not recognize the constraint set C, WeTune concretizes the q_{src} and q_{dest} according to the constraint C with the following three steps: First, we assign names to each symbol in q_{src} and q_{dest} according to those equivalence constraints including *RelEq*, *AttrsEq*, *PredEq* and *AggrEq*. Specifically, it puts the equivalent symbols into the same set, and all symbols in the same set will share a randomly generated name. For example, in Figure 2, $t2$, $t2'$ and $t4$ could be assigned with the name "T2". $c0$, $c0'$ and $c1$ could be assigned with the name $C1$. Second, for each attribute, we find the relation it belongs to according to the *SubAttrs* constraints. If an attribute with the name c belongs to a relation with the name t, then we change the attribute name from c to $t.c$. For the example in Figure 2, the name of attribute list $c0$ will be changed to $T1.C1$. Third, we construct the schema definition according to the attributes of relations. In Figure 2, T1's schema has 1 column $C1$.

Features	SPES	Built-in
Aggregation	✓	✗
UNION	✓	✗
NULL	✓	✓
OUTER JOIN	✓	✓
Complex Predicate	✓	✗
Predicate with NOT/XOR/OR	✓	✗
Integrity Constraint	✗	✓
Different # of input tables	✗	✓

Table 6: Comparison of the capabilities of SPES and WeTune's built-in verifier. ✓ indicates a feature is supported or partially supported. Complex predicates refer to predicates with arithmetic operations and CASE.

Table 6 compares different features supported by SPES and the built-in verifier. Compared with SPES, the built-in verifier does not support Aggregate and predicate with NOT/XOR/OR due to our implementation restriction. It cannot support UNION because the U-expression of UNION with Projection, which is in the form "$\sum + \sum$", cannot be converted into the FOL formula. Similarly, the other set operators, such as INTERSECT and DIFFERENCE, also cannot be supported by the built-in verifier. It is unnecessary for the built-in verifier to support complex predicates, because its enumerated query templates do not have concrete predicates. Compared with the built-in verifier, SPES cannot handle SQL query with integrity constraints. Furthermore, SPES cannot prove the equivalence of two queries if they have different input tables, as these queries can not be normalized to the same algebraic representation which is necessary for the proof. For example, SPES cannot prove the equivalence between "SELECT DISTINCT T.* FROM T" with "SELECT DISTINCT T.* FROM T LEFT JOIN R On T.k = R.k", which can be proved by the built-in verifier. However, WeTune does not fully utilize SPES because the current rule enumerator can not enumerate plan templates having concrete aggregation functions, complex predicates and predicates connected by XOR, OR and NOT. This is considered as future work. A detailed comparison between the built-in verifier and SPES can be found in Section 8.5 and Table 7.

6 Selecting Useful Rules

After generating the promising rules, WeTune empirically evaluates their usefulness. The basic idea is to collect queries from real-world applications and evaluate which rules can rewrite those queries into a more efficient form. Ideally, rewrites should be done by the database optimizer using existing rewriting techniques. However, to work with non-open-source databases, WeTune performs rewrites outside of the database.

$$q_{src} : Proj_{a_0}(LJoin_{a_1,a_2}(Input_{r_0}, Input_{r_1}))$$
$$C = \{SubAttrs(a_0, a_{r_0}), SubAttrs(a_1, a_{r_0}),$$
$$SubAttrs(a_2, a_{r_1}), Unique(r_1, a_2), \cdots\}$$

$\hat{q}$:
```
SELECT T.a AS k FROM T LEFT JOIN S ON T.b=S.c
integrity constraint: S.c is unique key
```

Figure 7: An example of generated probing query $\hat{q}$.

WeTune's rewriting logic is based on simple greedy search. Given a query, it iteratively applies the rule that results in the most simplified target query (aka one with the fewest relational operators of each type). There can be more than one such rule at each iterative step. The iterative process terminates when no rewrites are possible. WeTune then obtains the cost estimate of each rewritten query from the existing database using the database's cost estimator, e.g., MySQL supports retrieving estimated cost by EXPLAIN EXTENDED command, MS SQL Server supports the same function by turning on SHOW_PLAN_ALL option. WeTune measures the actual performance of the most cost-efficient version of the query. To run the query and its rewritten version, we populate the database tables according to the schema and integrity constraints using randomly generated data. If the performance is improved by rewriting, then the corresponding rewrite rules are considered useful.

7 Additional Optimization

WeTune proposes two extra optimization strategies to reduce the redundant rules and eliminate ORDER BY in SQL statements.

Reducing redundant rules. Multiple rules can be composed to rewrite a query. For example, consider a query q and three rules R_1, R_2, R_3, we may get the same query p after rewriting q by (1) consecutively applying R_1, R_2 or (2) applying R_3. Thus, R_3 is redundant and can be replaced by the composite of R_1 and R_2. During rule discovery, it is desirable to reduce such redundant rules. Formally, given a set of rules $\mathbb{R}$ and a rule $R \in \mathbb{R}$, R is *reducible* under $\mathbb{R}$ if

$$\forall q.(Rewrite(\mathbb{R}, q) = Rewrite(\mathbb{R} - \{R\}, q))$$

It is impossible to check all queries. Instead, WeTune generates a concrete probing query $\hat{q}$ and concrete constraints according to R's source plan template and constraints. First, WeTune concretizes q_{src} to be $\hat{q}$ according to the steps of concretizing q_{src} for SPES (Section 5.2). Second, WeTune adds concrete integrity constraints for $\hat{q}$ according to *NotNull*, *Unique* and *RefAttrs* constraints in the rule. Figure 7 shows an example. $\hat{q}$ must be the minimal pattern that R is applicable to. i.e., any query that R is applicable to must contain the pattern $\hat{q}$. Thus, to decide the reducibility, it is sufficient to check whether the following condition is true: $Rewrite(\mathbb{R}, \hat{q}) = Rewrite(\mathbb{R} - \{R\}, \hat{q})$

Eliminating ORDER BY. Although WeTune does not support ORDER BY, WeTune can remove it from the query when it does not affect the query semantic. This is based on the insight that in SQL, an "ORDER BY" operator in the subquery can be useless when the outer query does not perform computations that can be affected by the order of tuples in the subquery result (e.g., aggregates a constant value from the subquery). In such cases, WeTune will directly eliminate ORDER BY in the statements.

8 Evaluation

The evaluation aims to answer the following questions:

Q1. How many new useful rules can WeTune discover?

Q2. How many new queries can WeTune optimize over existing systems for real-world applications?

Q3. How does WeTune's built-in verifier compare with SPES?

8.1 Experimental Setup

Implementation. We have built WeTune from scratch, which has about 40k lines of Java code. It takes the max plan template size as the parameter and outputs a set of non-reducible and promising rules. WeTune can also automatically check the usefulness of these rules by cooperating with existing database systems, including MySQL, PostgreSQL and MS SQL Server. Thus, besides the rule enumerator and verifier, it also contains a SQL parser, a query plan builder and a benchmark framework that evaluates SQL performance. Specifically, the built-in verifier is based on the SMT solver Z3 [12].

Generating Rules. WeTune enumerates all query plan templates up to size 4, yielding 3113 distinct templates. WeTune finds 1106 promising and non-reducible rules in 36 hours (on 120 CPU cores in total), among which 32 hours were spent in verification. Each potential rule takes about 50 ms on average to verify. For each rule, WeTune invokes the SMT solver 383 times on average to search for the most relaxed constraint set.

Workload. We use two workloads for the evaluation: one is a real-world workload, another is the Apache Calcite test suite [5]. For the real-world workload, we collect SQL queries from 20 open-source web applications on GitHub with the most stars for evaluation (the full list is included in our extended version [49]). These applications come from various genres, such as e-commerce, content management, discussion forum and social network. The number of contributors varies from 1 (1,902 stars) to 2,007 (22,203 stars). We collected 8,518 unique queries by running unit tests bundled with the source code. The Calcite test suite comprises 232 pairs of queries (464 individual queries) that are known equivalent, and all these queries can be rewritten by the rules in Calcite.

Evaluating Rules. When selecting useful rules, all rules are evaluated based on MS SQL Server 2019. The queries used to evaluate rules include both workloads described above. When executing queries on the database to evaluate latency, we populate four different tables. Two tables have 1K rows, while the other two have 100K rows. For every two tables with the same number of rows, one of them is populated with random data generated according to the uniform distribution, while the other one is populated with random data generated according to the Zipfian distribution with a skewed parameter of 1.25.

Testbed. All experiments are run on a server with a 20-core (2 sockets) Intel E5-2650 v3 CPU, 126 GB DRAM, and 1 TB SSD. The end-to-end latency of every query is evaluated on MS SQL Server 2019. We implement a dedicated client program that issues database queries and resides on the same machine as the database to eliminate network communication overhead. For a given query, the client randomizes the parameters in the query with extra care to avoid that every execution always directly fetches results from the database cache and to prevent the output result set is always

empty. Each query is repeatedly executed 200 times in a closed loop (the first 100 times serve as warmup and are not counted into the result). When comparing WeTune with the rewriter in existing databases, we use the rewriter in Microsoft SQL Server 2019.

8.2 New Rewrite Rules

Table 7 shows the rules found by WeTune. There are 35 distinct rules which are useful for the evaluated queries. Among these rules, 9 rules are missing in MS SQL Server, 22 are missing in Calcite, and 5 are missing in both systems. 34 rules are discovered with the 8518 queries collected from the web application; only rule 35 is discovered with the queries in the Calcite test suite.

For the used verifier, 15 rules can be proved by both the built-in verifier and SPES. 16 rules can only be proved by the built-in verifier. SPES fails to prove these rules because 10 cases involve integrity constraints, 4 have mismatched input tables between the q_{src} and q_{dest}, and 2 cases (Rule 19 & 21) are related to SPES implementation. Taking rule 19 as an example, SPES fails to prove its correctness because we replace the predicate symbols in the query templates with the user-defined function. However, SPES does not consider that two user-defined functions are equivalent even if they have the same function name [4]. Compared with the built-in verifier, 4 rules can only be proved by SPES, as these rules have certain features only supported by SPES.

8.3 Queries Optimized by WeTune

We try to use generated rules of WeTune to rewrite both the queries studied in Section 2.2 and collected from real-world applications to see how many of them can be optimized by our discovered rules but cannot be optimized by existing systems.

The number of queries rewritten. For the 50 issues we have studied, WeTune can optimize 76% (38) of them, while MS SQL Server and Calcite can only optimize 46% (23) and 8%(4) of them. WeTune is unable to rewrite the remaining 12 queries due to two reasons. First, 9 of them need to rewrite the predicate expression or add a new predicate that does not equal to predicates in the original query (e.g., rewrite the predicate from "id IS NULL" to " project_id IS NULL" [1]). It requires finer-grained modeling and reasoning of the predicate expressions. The rest 3 of the queries need explicitly model the semantics of operators that WeTune currently does not support, including Aggregate and GROUP BY.

For 8518 queries collected from 20 real-world applications, WeTune can successfully rewrite 674 queries, among which 247 queries SQL Server fails to optimize (the other 427 queries can be effectively optimized). We manually check the remaining 7844 queries to investigate why WeTune cannot rewrite them. The main reason is that most queries (4251) only consist of SELECT-clause and WHERE-clause, without JOIN, subquery, Aggregate or any other clause. Optimizing such queries usually requires transformation at the physical execution level (e.g., index choice), which is beyond the scope of WeTune. The result shows WeTune can optimize more queries over existing databases.

We try to rewrite all 464 queries in the Calcite test suite. WeTune can rewrite 120 queries, among which 26 cannot be effectively optimized by MS SQL Server. For 23 queries of them, the rewriting performed by WeTune can achieve a 23.8% - 95.2% latency reduction than the rewriting of Calcite itself.

Latency reduction. To show the effectiveness of the optimizations found by WeTune, we compare the latency of the rewritten query with the original one on the same database for each of 273 queries that cannot be optimized by MS SQL Server (247 from applications plus 26 from Calcite test suite).

To know whether these rewrites are specific to certain workloads, we synthesized four workloads with varying table sizes (number of rows in the table) and data distribution, as summarized below:

	# of rows=10K	# of rows=1M
uniform dist.	workload A	workload B
zipfian dist. (θ = 1.5)	workload C	workload D

We implement a data generator inserting randomly generated rows into tables, which carefully maintains integrity constraints.

For workload A, WeTune can optimize at least 50% of the queries with more than 10% latency reduction and 17%, 18%, 30% reduction for workload B, C and D, respectively. WeTune can also optimize 13%-21% queries with at least a 90% latency reduction for all four workloads. This demonstrates that the rewrites are not specific to a certain table size or data distribution.

8.4 Case Study

Take the second query in Table 1 as an example of finding sequences of useful rules to optimize a query. First, WeTune iteratively generates new queries via rewrite rules, which takes 1.5s. Second, it consumes 5.3s to use the cost estimator in MS SQL Server to evaluate generated new queries. Then, we evaluate the end-to-end latency of every generated query by issuing it to the database. This step takes 12s, which indicates that we can find the sequence of rules that can produce the query with better performance within a reasonable amount of time.

Figure 8 shows each step of the best sequence of rewrite rules for the example above. First, the IN-selection is transformed to INNER JOIN in (2). Then, the predicate below the INNER JOIN is pulled up above it in (3). Usually, pushing down predicate below a JOIN is a standard optimization technique that can eagerly reduce the number of rows. However, in this case, pulling the predicate up enables new optimization opportunities that lead to a more efficient query. Next, the column "m.commit_id" used in the predicate is replaced by "n.commit_id" in (4). This replacement is guaranteed correct because the ON-condition "n.id=m.id" and the uniqueness property of primary key collectively imply that "m.commit_id=n.commit_id" holds for each row in the result set of the JOIN. Last, the table source t1 is eliminated by applying the JOIN-elimination rule.

Some rules that rewrite the source query plan to a similar plan are still useful, such as rule 17 and rule 18. For example, the query "Select T.y From R Inner Join T On R.x=T.y" will become "Select R.x From R Inner Join T On R.x=T.y" after applying rule 17. This rewrite allows WeTune to further apply rule 7 to eliminate the join when rule 7's constraints are met. Similarly, Rule 18 is useful because it might enable the subsequent application of rule 8.

8.5 Built-in Verifier vs. SPES

SPES is the state-of-the-art SQL equivalence verifier [50]. We try to compare the built-in verifier with SPES via two workloads: one

[4] A hypothesis of such design is SPES may aim to support some functions like RANDOM, which is not considered by WeTune.

No.	Source Plan Template	Destination Plan Template	Extra Constraints	Verifier	Calcite	MS
1	$\text{Sel}_{\text{p,r.a}_0}(\text{Proj}_{\text{r.a}_1}(\text{r}))$	$\text{Proj}_{\text{r.a}_1}(\text{Sel}_{\text{p,r.a}_0}(\text{r}))$	SubAttrs(a_1, a_0)	B	Y	Y
2	$\text{Dedup}(\text{Proj}_{\text{r.a}}(\text{r}))$	$\text{Proj}_{\text{r.a}}(\text{r})$	Unique(r, a)	W	N	Y
3	$\text{Sel}_{\text{p,r.a}}(\text{Sel}_{\text{p,r.a}}(\text{r}))$	$\text{Sel}_{\text{p,r.a}}(\text{r})$		B	Y	Y
4	$\text{InSub}_{\text{r}_0.\text{a}_0}(\text{InSub}_{\text{r}_0.\text{a}_0}(\text{r}_0, \text{r}_1), \text{r}_1)$	$\text{InSub}_{\text{r}_0.\text{a}_0}(\text{r}_0, \text{r}_1)$		W	N	N
5	$\text{Proj}_{\text{r.a}_0}(\text{Sel}_{\text{p,r.a}_1}(\text{Proj}_{\text{r.a}_2}(\text{r})))$	$\text{Proj}_{\text{r.a}_0}(\text{Sel}_{\text{p,r.a}_1}(\text{r}))$	SubAttrs(a_0, a_2),SubAttrs(a_1, a_2)	B	Y	Y
6	$\text{LJoin}_{\text{r}_0.\text{a}_0,\text{r}_1.\text{a}_1}(\text{r}_0, \text{r}_1)$	$\text{IJoin}_{\text{r}_0.\text{a}_0,\text{r}_1.\text{a}_1}(\text{r}_0, \text{r}_1)$	RefAttrs(r_0, a_0, r_1, a_1),NotNull(r_0, a_0)	W	N	Y
7	$\text{Proj}_{\text{r}_0.\text{a}_2}(\text{IJoin}_{\text{r}_0.\text{a}_0,\text{r}_1.\text{a}_1}(\text{r}_0, \text{r}_1))$	$\text{Proj}_{\text{r}_0.\text{a}_2}(\text{r}_0)$	RefAttrs(r_0, a_0, r_1, a_1),NotNull(r_0, a_0),Unique(r_1, a_1)	W	N	Y
8	$\text{Proj}_{\text{r}_0.\text{a}_2}(\text{Sel}_{\text{p,r}_0.\text{a}_3}(\text{IJoin}_{\text{r}_0.\text{a}_0,\text{r}_1.\text{a}_1}(\text{r}_0, \text{r}_1)))$	$\text{Proj}_{\text{r}_0.\text{a}_2}(\text{Sel}_{\text{p,r}_0.\text{a}_3}(\text{r}_0))$	RefAttrs(r_0, a_0, r_1, a_1),NotNull(r_0, a_0),Unique(r_1, a_1)	W	N	C
9	$\text{Dedup}(\text{Proj}_{\text{r}_0.\text{a}_2}(\text{IJoin}_{\text{r}_0.\text{a}_0,\text{r}_1.\text{a}_1}(\text{r}_0, \text{r}_1)))$	$\text{Dedup}(\text{Proj}_{\text{r}_0.\text{a}_2}(\text{r}_0))$	RefAttrs(r_0, a_0, r_1, a_1),NotNull(r_0, a_0)	W	N	Y
10	$\text{Dedup}(\text{Proj}_{\text{r}_0.\text{a}_2}(\text{Sel}_{\text{p,r}_0.\text{a}_3}(\text{IJoin}_{\text{r}_0.\text{a}_0,\text{r}_1.\text{a}_1}(\text{r}_0, \text{r}_1))))$	$\text{Dedup}(\text{Proj}_{\text{r}_0.\text{a}_2}(\text{Sel}_{\text{p,r}_0.\text{a}_3}(\text{r}_0)))$	RefAttrs(r_0, a_0, r_1, a_1),NotNull(t_0, a_0)	W	N	C
11	$\text{Proj}_{\text{r}_0.\text{a}_2}(\text{LJoin}_{\text{r}_0.\text{a}_0,\text{r}_1.\text{a}_1}(\text{r}_0, \text{r}_1))$	$\text{Proj}_{\text{r}_0.\text{a}_2}(\text{a}_0)$	Unique(r_1, a_1)	W	N	Y
12	$\text{Proj}_{\text{r}_0.\text{a}_3}(\text{Sel}_{\text{p,r}_0.\text{a}_2}(\text{LJoin}_{\text{r}_0.\text{a}_0,\text{r}_1.\text{a}_1}(\text{r}_0, \text{r}_1)))$	$\text{Proj}_{\text{r}_0.\text{a}_3}(\text{Sel}_{\text{p,r}_0.\text{a}_2}(\text{r}_0))$	Unique(r1, a1)	W	N	Y
13	$\text{Dedup}(\text{Proj}_{\text{r}_0.\text{a}_2}(\text{LJoin}_{\text{r}_0.\text{a}_0,\text{t}_1.\text{r}_2}(\text{r}_0, \text{r}_1)))$	$\text{Dedup}(\text{Proj}_{\text{r}_0.\text{a}_2}(\text{a}_0))$		W	N	Y
14	$\text{Dedup}(\text{Proj}_{\text{r}_0.\text{a}_3}(\text{Sel}_{\text{p,r}_0.\text{a}_2}(\text{LJoin}_{\text{r}_0.\text{a}_0,\text{r}_1.\text{a}_1}(\text{r}_0, \text{r}_1))))$	$\text{Dedup}(\text{Proj}_{\text{r}_0.\text{a}_3}(\text{Sel}_{\text{p,r}_0.\text{a}_2}(\text{r}_0)))$		W	N	Y
15	$\text{InSub}_{\text{r.a}}(\text{r}, \text{Proj}_{\text{r}'.\text{a}}(\text{r}'))$	r	NotNull(r, a)	W	Y	N
16	$\text{Proj}_{\text{r.a}}(\text{IJoin}_{\text{r.a,r}'.\text{a}}(\text{r}, \text{r}'))$	$\text{Proj}_{\text{r.a}}(\text{r})$	NotNull(r, a), Unique(r, a)	W	N	N
17	$\text{Proj}_{\text{r}_1.\text{a}_1}(\text{IJoin}_{\text{r}_0.\text{a}_0,\text{r}_1.\text{a}_1}(\text{r}_0, \text{r}_1))$	$\text{Proj}_{\text{r}_0.\text{a}_0}(\text{IJoin}_{\text{r}_0.\text{a}_0,\text{r}_1.\text{a}_1}(\text{r}_0, \text{r}_1))$		B	N	N
18	$\text{Proj}_{\text{r}_1.\text{a}_1}(\text{Sel}_{\text{p,r}_0.\text{a}_2}(\text{IJoin}_{\text{r}_0.\text{a}_0,\text{r}_1.\text{a}_1}(\text{r}_0, \text{r}_1)))$	$\text{Proj}_{\text{r}_0.\text{a}_0}(\text{Sel}_{\text{p,r}_0.\text{a}_2}(\text{IJoin}_{\text{r}_0.\text{a}_0,\text{r}_1.\text{a}_1}(\text{r}_0, \text{r}_1)))$		B	N	N
19	$\text{Sel}_{\text{p,r}_1.\text{a}_1}(\text{IJoin}_{\text{r}_0.\text{a}_0,\text{r}_1.\text{a}_1}(\text{r}_0, \text{r}_1))$	$\text{Sel}_{\text{p,r}_0.\text{a}_0}(\text{IJoin}_{\text{r}_0.\text{a}_0,\text{r}_1.\text{a}_1}(\text{r}_0, \text{r}_1))$		W	N	Y
20	$\text{IJoin}_{\text{r}_1.\text{a}_1,\text{r}_2.\text{a}_2}(\text{IJoin}_{\text{r}_0.\text{a}_0,\text{r}_1.\text{a}_1}(\text{r}_0, \text{r}_1), \text{r}_2)$	$\text{IJoin}_{\text{r}_0.\text{a}_0,\text{r}_2.\text{a}_2}(\text{IJoin}_{\text{r}_0.\text{a}_0,\text{r}_1.\text{a}_1}(\text{r}_0, \text{r}_1), \text{r}_2)$		B	N	Y
21	$\text{LJoin}_{\text{r}_1.\text{a}_1,\text{r}_2.\text{a}_2}(\text{IJoin}_{\text{r}_0.\text{a}_0,\text{r}_1.\text{a}_1}(\text{r}_0, \text{r}_1), \text{r}_2)$	$\text{LJoin}_{\text{r}_0.\text{a}_0,\text{r}_2.\text{a}_2}(\text{IJoin}_{\text{r}_0.\text{a}_0,\text{r}_1.\text{a}_1}(\text{r}_0, \text{r}_1), \text{r}_2)$		W	N	Y
22	$\text{Proj}_{\text{r}_0.\text{a}_2}(\text{IJoin}_{\text{r}_0.\text{a}_0,\text{r}_1.\text{a}_1}(\text{r}_0, \text{r}_1))$	$\text{Proj}_{\text{r}_0.\text{a}_2}(\text{IJoin}_{\text{r}_1.\text{a}_1,\text{r}_0.\text{a}_0}(\text{r}_1, \text{r}_0))$		B	Y	Y
23	$\text{IJoin}_{\text{r}_0.\text{a}_0,\text{r}_1.\text{a}_1}(\text{r}_0, \text{IJoin}_{\text{r}_1.\text{a}_2,\text{r}_2.\text{a}_3}(\text{r}_1, \text{r}_2))$	$\text{IJoin}_{\text{r}_1.\text{a}_2,\text{r}_2.\text{a}_3}(\text{IJoin}_{\text{r}_0.\text{a}_0,\text{r}_1.\text{a}_1}(\text{r}_0, \text{r}_1), \text{r}_2)$		B	Y	Y
24	$\text{Proj}_{\text{r}_0.\text{a}_2}(\text{InSub}_{\text{r}_0.\text{a}_0}(\text{r}_0, \text{Proj}_{\text{r}_1.\text{a}_1}(\text{r}_1)))$	$\text{Proj}_{\text{r}_0.\text{a}_2}(\text{IJoin}_{\text{r}_0.\text{a}_0,\text{r}_1.\text{a}_1}(\text{r}_0, \text{r}_1))$	Unique(r_1, a_1)	B	Y	Y
25	$\text{Proj}_{\text{r}_0.\text{a}_2}(\text{IJoin}_{\text{r}_0.\text{a}_0,\text{r}_1.\text{a}_1}(\text{r}_0, \text{Dedup}(\text{Proj}_{\text{r}_1.\text{a}_1}(\text{r}_1))))$	$\text{Proj}_{\text{r}_0.\text{a}_2}(\text{InSub}_{\text{r}_0.\text{a}_0}(\text{r}_0, \text{Proj}_{\text{r}_1.\text{a}_1}(\text{r}_1)))$		B	N	Y
26	$\text{Dedup}(\text{Proj}_{\text{r}_0.\text{a}_2}(\text{IJoin}_{\text{r}_0.\text{a}_0,\text{r}_1.\text{a}_1}(\text{r}_0, \text{Dedup}(\text{r}_1))))$	$\text{Dedup}(\text{Proj}_{\text{r}_0.\text{a}_2}(\text{IJoin}_{\text{r}_0.\text{a}_0,\text{r}_1.\text{a}_1}(\text{r}_0, \text{r}_1)))$		W	N	Y
27	$\text{IJoin}_{\text{r}_0.\text{a}_0,\text{r}_1.\text{a}_1}(\text{r}_0, \text{Sel}_{\text{p,r}_1.\text{a}_2}(\text{r}_1))$	$\text{Sel}_{\text{p,r}_1.\text{a}_2}(\text{IJoin}_{\text{r}_0.\text{a}_0,\text{r}_1.\text{a}_1}(\text{r}_0, \text{r}_1))$		B	Y	Y
28	$\text{Sel}_{\text{p,r}_1.\text{a}_2}(\text{IJoin}_{\text{r}_0.\text{a}_0,\text{r}_1.\text{a}_1}(\text{r}_0, \text{r}_1))$	$\text{IJoin}_{\text{r}_0.\text{a}_0,\text{r}_1.\text{a}_1}(\text{r}_0, \text{Sel}_{\text{p,r}_1.\text{a}_2}(\text{r}_1))$		B	Y	Y
29	$\text{Proj}_{\text{r}_0.\text{a}_2}(\text{IJoin}_{\text{r}_0.\text{a}_0,\text{r}_1.\text{a}_1}(\text{r}_0, \text{Proj}_{\text{r}_1.\text{a}_1}(\text{r}_1)))$	$\text{Proj}_{\text{r}_0.\text{a}_2}(\text{IJoin}_{\text{r}_0.\text{a}_0,\text{r}_1.\text{a}_1}(\text{r}_0, \text{r}_1)))$		B	N	Y
30	$\text{Sel}_{\text{p,r.a}_0}(\text{IJoin}_{\text{r.a}_1,\text{r}'.\text{a}_1}(\text{r}, \text{r}'))$	$\text{Sel}_{\text{p,r}'.\text{a}_0}(\text{IJoin}_{\text{r.a}_1,\text{r}'.\text{a}_1}(\text{r}, \text{r}'))$	Unique(r, a_1)	B	N	N
31	$\text{Proj}_{\text{r}_0.\text{a}_0}(\text{LJoin}_{\text{r}_0.\text{a}_1,\text{r}_1.\text{a}_2}(\text{Proj}_{\text{r}_0.\text{a}_3}(\text{r}_0), \text{r}_1))$	$\text{Proj}_{\text{r}_0.\text{a}_0}(\text{LJoin}_{\text{r}_0.\text{a}_1,\text{r}_1.\text{a}_2}(\text{r}_0, \text{r}_1))$		B	Y	Y
32	$\text{Proj}_{\text{r}_0.\text{a}_0}(\text{LJoin}_{\text{r}_0.\text{a}_1,\text{r}_1.\text{a}_2}(\text{r}_0, \text{Proj}_{\text{r}_1.\text{a}_3}(\text{r}_1)))$	$\text{Proj}_{\text{r}_0.\text{a}_0}(\text{LJoin}_{\text{r}_0.\text{a}_1,\text{r}_1.\text{a}_2}(\text{r}_0, \text{r}_1))$		S	Y	Y
33	$\text{Agg}_{\text{r.a}_0,\text{f,r.a}_1,\text{p}_0}(\text{Filter}_{\text{p}_1,\text{a}_2}(\text{Proj}_{\text{r.a}_3}(\text{r})))$	$\text{Agg}_{\text{r.a}_0,\text{f,r.a}_1,\text{p}_0}(\text{Filter}_{\text{p}_1,\text{a}_2}(\text{r}))$	SubAttrs(a_0, a_3), SubAttrs(a_1, a_3), SubAttrs(a_2, a_3)	S	Y	Y
34	$\text{Agg}_{\text{r}_0.\text{a}_0,\text{f,r}_0.\text{a}_1,\text{p}}(\text{IJoin}_{\text{r}_0.\text{a}_2,\text{r}_1.\text{a}_3}(\text{Proj}_{\text{r.a}_4}(\text{r}_0), \text{r}_1))$	$\text{Agg}_{\text{r}_0.\text{a}_0,\text{f,r}_0.\text{a}_1,\text{p}}(\text{IJoin}_{\text{r}_0.\text{a}_2,\text{r}_1.\text{a}_3}(\text{r}_0, \text{r}_1))$	SubAttrs(a_0, a_4), SubAttrs(a_1, a_4), SubAttrs(a_2, a_4)	S	N	Y
35	$\text{Agg}_{\text{r.a}_0,\text{f,r.a}_1,\text{p}_0}(\text{Filter}_{\text{p}_0,\text{r.a}_0}(\text{r}))$	$\text{Agg}_{\text{r.a}_0,\text{f,r.a}_1,\text{p}_0}(\text{r})$		S	Y	N

Table 7: Useful rewrite rules found by WeTune. The Verifier column indicates which verifier can prove the rule. *W* means the built-in verifier, *S* means SPES, and *B* means both. The Calcite and MS columns indicate whether Calcite and MS SQL Server support these rules. The tree structure of the plan template is flattened by pre-order traversal. Each r_i represents an input table. Each a_i represents an attribute list. Each p represents a predicate. IJoin **is the abbreviation for** InnerJoin**, and** LJoin **is for** LeftJoin**. Multiple occurrences of the same symbol (i.e.,** r_i, a_i, p**) depict the equivalence constraint. Each** $\text{r}_\text{i}.\text{a}_\text{j}$ **stands for a constraint** $\text{SubAttrs}(\text{a}_\text{j}, \text{a}_{\text{r}_\text{i}})$**. Other types of constraints are listed in the column Extra Constraints. For rule 15 and 16,** r **and** r′ **denote two distinct occurrences of the same relation (e.g., "SELECT r.* FROM tbl AS r INNER JOIN tbl AS r' ON r.k=r'.k"). For rule 8 and 10, SQL Server can conditionally (C) eliminate the JOIN only if the attributes** a_3 **is different from** a_0**.**

is the 861 rules generated using the built-in verifier. Another is the 232 pairs of equivalent SQL in Calcite test suite.

Rules generated by the built-in verifier. With the built-in verifier, WeTune is able to enumerate 861 promising and non-reducible rules. Among these rules, SPES successfully verifies 41 rules. Among the 820 that are not verified, 725 are due to that SPES's current implementation does not support integrity constraints and 95 are due to mismatched numbers of input tables on both sides.

Calcite Test Suite. The Calcite test suite comprises 232 pairs of queries. Each pair includes two equivalent SQL. SPES can successfully verify the equivalence of 95 query pairs of them, while the built-in verifier can prove the equivalence of 73 query pairs. Specifically, 55 pairs can be proved by both the built-in verifier and SPES. The number of pairs that the built-in verifier can prove is less than that can be proved by SPES because most rewrite rules in the test suite involve unsupported features of the built-in verifier, such as complex predicate. However, these features are supported by SPES.

9 Related Work

Query equivalence verification. Recently, researchers have proposed several systems [8–10, 51] to prove the equivalence of SQL queries formally. There are two approaches: some are based on proof assistants [8–10] while others are based on SMT [50, 51]. For the former approach, the state-of-the-art checker [8] uses an algebraic approach to verify the correctness of rules. Although their algebraic approach can model complex query structures based on the bag semantics, it lacks support for three-value-logic reasoning. For the SMT-based approach, recent work [50, 51] proposed *symbolic representation* of the query and leveraged the SMT solver to efficiently prove the equivalence of queries. But these systems lack the support of integrity constraints. WeTune overcomes some of their disadvantages by extending the algebraic approach with three-value-logic reasoning and supporting features such as integrity constraints.

Superoptimization. As a compiler optimization technique, superoptimization [2, 34] aims to find the optimal code sequence of a

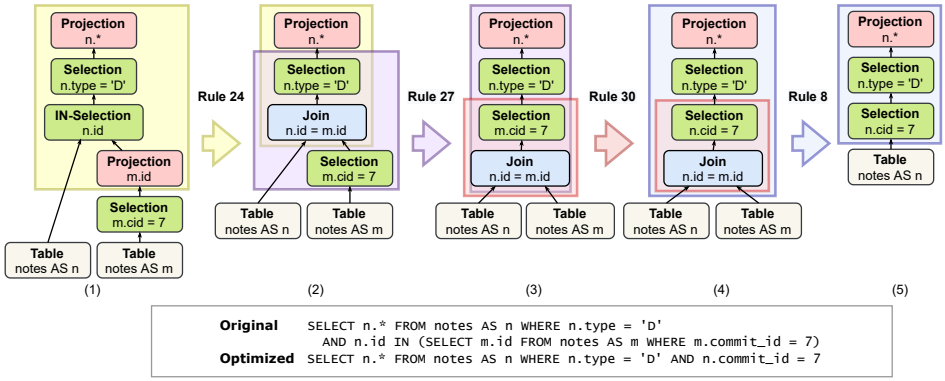


Figure 8: An example sequence of rules discovered by WeTune to optimize a SQL query of a real-world query. The notes.id is the primary key of table notes. The cid in the figure is the abbreviation of commit_id. Each colored arrow in the figure represents a rule with the rewrite rule index above the arrow. The sub-plan in the same color is the source and destination plan template of the rule. The corresponding constraints are omitted.

set of instructions, which inspires WeTune. TASO [23] leverages superoptimization to find rewrite rules to rewrite deep neural networks (DNN). However, these works target different scenarios from query rewrite. WeTune needs to adopt different enumeration and verification methods. For example, in terms of enumeration, WeTune considers the relations of symbols (constraint) in SQL rewrite rules and enumerates all possible constraints. DNN operators have simpler parameters, and TASO only considers the relation between input/output operators. Ruler [38] has proposed a framework that abstracts the "search + verification" methods based on equality saturation to reduce candidate generation and selection cost. It is a general approach instead of specifically targeting SQL queries. WeTune could potentially use this framework to further improve the speed of discovery.

Query optimization. There has been a long line of work for query optimization, roughly divided into two categories depending on the search strategy of query plans. One is through a stratified approach [29, 42, 45], which first rewrites the logical query plan using transformation rules and then performs a cost-based search to map the logical plan to a physical plan. The other is through a unified approach [3, 14, 16, 18, 37], which unifies the logical to logical and logical to physical transformation into one stage. Recently, there has been a trend in adopting deep learning to query optimization [24, 26, 30–32, 47]. Given a set of rewrite rules, LearnedRewrite [52] is able to find the optimal rewrite order by using Monte Carlo Tree with learned cost models. However, these methods require manually written transformation rules and are orthogonal with the goal of WeTune.

10 Limitations

WeTune has the following two major limitations.

Incompleteness. One limitation is the incompleteness of the built-in verifier. First, due to the unbounded nature of the $\sum$ operator, U-expression fundamentally exceeds the expressiveness of FOL. Currently, only cases listed in table 5 can be translated to FOL and automatically verified by the SMT solver. How to automatically transform any U-expression into FOL formulas is left as future work. Second, the translated formula does not always fall into a decidable fragment of the SMT solver; thus may lead to timeout and consequently miss useful rules.

Unsupported SQL features. Another limitation is that the built-in verifier currently only supports rules containing operators listed in table 2. Furthermore, WeTune does not support recursive queries. As described in Section 5.2, some features are unsupported, such as UNION. Some features are partially supported, such as NULL. As described in Section 5.1.1, WeTune currently only considers the impact of NULL on operators in Table 3. Supporting more features is left as future work. Although some SQL features are unsupported, the soundness of WeTune holds for non-recursive queries. In other words, rewriting a query plan with rules obtained by WeTune can guarantee equivalent semantics. This is because, for every non-recursive query, even if it contains unsupported features, replacing its sub-plan without such features with another equivalent plan will not alter its original semantics.

11 Conclusion

This paper presents WeTune, which can automatically discover the rewrite rules for SQL queries. It enumerates all valid logical query plans up to a certain size to discover equivalent plans based on a new SMT-based verifier. We apply the rules discovered by WeTune on SQL queries collected from the 20 most popular open-source web applications on GitHub. WeTune successfully optimizes 247 queries that existing databases cannot optimize, resulting in substantial performance improvements.

Acknowledgments

We thank all anonymous reviewers for their constructive feedback and suggestions. This work is supported in part by National Natural Science Foundation of China (No. 62132014, 61902242, 62172272), the HighTech Support Program from Shanghai Committee of Science and Technology (No. 20ZR1428100). Ding Ding is supported by a DeepMind fellowship. Zhaoguo Wang (zhaoguowang@sjtu.edu.cn) is the corresponding author.

References

[1] Douglas Barbosa Alexandre. 2018. Improve the query performance to find unverified projects. https://gitlab.com/gitlab-org/gitlab/-/commit/11e93a9a4c2ac1b5bd4d32a93a949fc8afbcc449?merge_request_iid=5348.

[2] Sorav Bansal and Alex Aiken. 2006. Automatic generation of peephole superoptimizers. *ACM SIGARCH Computer Architecture News* 34, 5 (2006), 394–403.

[3] Edmon Begoli, Jesús Camacho-Rodríguez, Julian Hyde, Michael J. Mior, and Daniel Lemire. 2018. Apache Calcite: A Foundational Framework for Optimized Query Processing Over Heterogeneous Data Sources. In *Proceedings of the 2018 International Conference on Management of Data* (Houston, TX, USA) *(SIGMOD '18)*. Association for Computing Machinery, New York, NY, USA, 221–230. https://doi.org/10.1145/3183713.3190662

[4] Andreas Brandl. 2018. Replace OR clause with UNION. https://gitlab.com/gitlab-org/gitlab-foss/-/merge_requests/17088#note_59749778

[5] Apache Calcite. 2021. Calcite Test Suite. https://github.com/georgia-tech-db/spes/blob/main/testData/calcite_tests.json.

[6] Jack Chen, Samir Jindel, Robert Walzer, Rajkumar Sen, Nika Jimsheleishvilli, and Michael Andrews. 2016. The MemSQL Query Optimizer: A modern optimizer for real-time analytics in a distributed database. *Proceedings of the VLDB Endowment* 9, 13 (2016), 1401–1412.

[7] Hugh Darwen Chris J Date. 1996. *A Guide to the SQL Standard, Forth Edition*. Addison-Wesley Professional. https://www.amazon.com/Guide-SQL-Standard-4th/dp/0201964260

[8] Shumo Chu, Brendan Murphy, Jared Roesch, Alvin Cheung, and Dan Suciu. 2018. Axiomatic Foundations and Algorithms for Deciding Semantic Equivalences of SQL Queries. *Proc. VLDB Endow.* 11, 11 (July 2018), 1482–1495. https://doi.org/10.14778/3236187.3236200

[9] Shumo Chu, Chenglong Wang, Konstantin Weitz, and Alvin Cheung. 2017. Cosette: An Automated Prover for SQL.. In *Proceedings of the 8th Biennial Conference on Innovative Data Systems Research* (Chaminade, California, USA) *(CIDR '17)*.

[10] Shumo Chu, Konstantin Weitz, Alvin Cheung, and Dan Suciu. 2017. HoTTSQL: Proving Query Rewrites with Univalent SQL Semantics. *SIGPLAN Not.* 52, 6 (June 2017), 510–524. https://doi.org/10.1145/3140587.3062348

[11] Spree Commerce. 2021. Spree. https://github.com/spree/spree.

[12] Leonardo De Moura and Nikolaj Bjørner. 2008. Z3: An Efficient SMT Solver. In *Proceedings of the Theory and Practice of Software, 14th International Conference on Tools and Algorithms for the Construction and Analysis of Systems* (Budapest, Hungary) *(TACAS '08/ETAPS'08)*. Springer-Verlag, Berlin, Heidelberg, 337–340.

[13] Diaspora. 2021. Diaspora. https://github.com/diaspora/diaspora.

[14] Visweswara Sai Prashanth Dintyala, Arpit Narechania, and Joy Arulraj. to appear. SQLCheck: Automated Detection and Diagnosis of SQL Anti-Patterns. (to appear).

[15] GitLab. 2021. GitLab. https://gitlab.com/gitlab-org/gitlab.

[16] Goetz Graefe. 1995. The cascades framework for query optimization. *IEEE Data Eng. Bull.* 18, 3 (1995), 19–29.

[17] Goetz Graefe and David J DeWitt. 1987. The EXODUS optimizer generator. In *Proceedings of the 1987 ACM SIGMOD international conference on Management of data*. 160–172.

[18] Goetz Graefe and William J McKenna. 1993. The volcano optimizer generator: Extensibility and efficient search. In *Proceedings of IEEE 9th International Conference on Data Engineering*. IEEE, 209–218.

[19] Todd J Green, Grigoris Karvounarakis, and Val Tannen. 2007. Provenance semirings. In *Proceedings of the twenty-sixth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*. 31–40.

[20] Paolo Guagliardo and Leonid Libkin. 2017. A formal semantics of SQL queries, its validation, and applications. *Proceedings of the VLDB Endowment* 11, 1 (2017), 27–39.

[21] Adam Hegyi. 2020. Suboptimal Query in Gitlab. https://gitlab.com/gitlab-org/gitlab/-/merge_requests/34364.

[22] ISO/IEC 9075-1:2003 2003. *Information technology – Database languages – SQL – Part 1: Framework (SQL/Framework)*. Standard. International Organization for Standardization, Geneva, CH.

[23] Zhihao Jia, Oded Padon, James Thomas, Todd Warszawski, Matei Zaharia, and Alex Aiken. 2019. TASO: Optimizing Deep Learning Computation with Automatic Generation of Graph Substitutions. In *Proceedings of the 27th ACM Symposium on Operating Systems Principles* (Huntsville, Ontario, Canada) *(SOSP '19)*. Association for Computing Machinery, New York, NY, USA, 47–62. https://doi.org/10.1145/3341301.3359630

[24] Andreas Kipf, Thomas Kipf, Bernhard Radke, Viktor Leis, Peter Boncz, and Alfons Kemper. 2019. Learned cardinalities: Estimating correlated joins with deep learning. In *Proceedings of the 9th Biennial Conference on Innovative Data Systems Research* (Asilomar, California, USA) *(CIDR '19)*.

[25] Civilized Discourse Construction Kit. 2021. Discourse. https://github.com/discourse/discourse.

[26] Sanjay Krishnan, Zongheng Yang, Ken Goldberg, Joseph Hellerstein, and Ion Stoica. 2018. Learning to optimize join queries with deep reinforcement learning. *arXiv preprint arXiv:1808.03196* (2018).

[27] Alon Y Levy, Inderpal Singh Mumick, and Yehoshua Sagiv. 1994. Query optimization by predicate move-around. In *VLDB*. 96–107.

[28] G. Linden. 2006. Marissa Mayer at Web 2.0. http://glinden.blogspot.com/2006/11/marissa-mayer-at-web-20.html/.

[29] Guy M Lohman. 1988. Grammar-like functional rules for representing query optimization alternatives. *ACM SIGMOD Record* 17, 3 (1988), 18–27.

[30] Ryan Marcus, Parimarjan Negi, Hongzi Mao, Chi Zhang, Mohammad Alizadeh, Tim Kraska, Olga Papaemmanouil, and Nesime Tatbul. 2019. Neo: A learned query optimizer. *arXiv preprint arXiv:1904.03711* (2019).

[31] Ryan Marcus and Olga Papaemmanouil. 2018. Deep reinforcement learning for join order enumeration. In *Proceedings of the First International Workshop on Exploiting Artificial Intelligence Techniques for Data Management*. 1–4.

[32] Ryan Marcus and Olga Papaemmanouil. 2019. Towards a Hands-Free Query Optimizer through Deep Learning. In *Proceedings of the 9th Biennial Conference on Innovative Data Systems Research* (Asilomar, California, USA) *(CIDR '19)*.

[33] Henry Massalin. 1987. Superoptimizer: A Look at the Smallest Program. In *Proceedings of the Second International Conference on Architectual Support for Programming Languages and Operating Systems* (Palo Alto, California, USA) *(ASPLOS II)*. IEEE Computer Society Press, Washington, DC, USA, 122–126. https://doi.org/10.1145/36206.36194

[34] Henry Massalin. 1987. Superoptimizer: a look at the smallest program. *ACM SIGARCH Computer Architecture News* 15, 5 (1987), 122–126.

[35] Sean McGivern. 2017. Speed up counting approvers when some are specified. https://gitlab.com/gitlab-org/gitlab/-/merge_requests/2196.

[36] Inderpal Singh Mumick, Sheldon J Finkelstein, Hamid Pirahesh, and Raghu Ramakrishnan. 1990. Magic is relevant. *ACM SIGMOD Record* 19, 2 (1990), 247–258.

[37] M. Muralikrishna. 1992. Improved Unnesting Algorithms for Join Aggregate SQL Queries. In *Proceedings of the 18th International Conference on Very Large Data Bases (VLDB '92)*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 91–102.

[38] Chandrakana Nandi, Max Willsey, Amy Zhu, Yisu Remy Wang, Brett Saiki, Adam Anderson, Adriana Schulz, Dan Grossman, and Zachary Tatlock. 2021. Rewrite Rule Inference Using Equality Saturation. *CoRR* abs/2108.10436 (2021). arXiv:2108.10436 https://arxiv.org/abs/2108.10436

[39] Nebulab. 2021. Solidus. https://github.com/solidusio/solidus.

[40] OpenProject. 2021. OpenProject. https://github.com/opf/openproject.

[41] Neil Patel. 2018. How Loading Time Affects Your Bottom Line. https://neilpatel.com/blog/speed-is-a-killer/.

[42] Hamid Pirahesh, Joseph M. Hellerstein, and Waqar Hasan. 1992. Extensible/Rule Based Query Rewrite Optimization in Starburst. (1992), 39–48. https://doi.org/10.1145/130283.130294

[43] Redmine. 2021. Redmine. https://github.com/redmine/redmine.

[44] Praveen Seshadri, Joseph M Hellerstein, Hamid Pirahesh, TY Cliff Leung, Raghu Ramakrishnan, Divesh Srivastava, Peter J Stuckey, and S Sudarshan. 1996. Cost-based optimization for magic: Algebra and implementation. In *Proceedings of the 1996 ACM SIGMOD international conference on Management of data*. 435–446.

[45] Praveen Seshadri, Joseph M. Hellerstein, Hamid Pirahesh, T. Y. Cliff Leung, Raghu Ramakrishnan, Divesh Srivastava, Peter J. Stuckey, and S. Sudarshan. 1996. Cost-Based Optimization for Magic: Algebra and Implementation. In *Proceedings of the 1996 ACM SIGMOD International Conference on Management of Data* (Montreal, Quebec, Canada) *(SIGMOD '96)*. Association for Computing Machinery, New York, NY, USA, 435–446. https://doi.org/10.1145/233269.233360

[46] Joshua Stein. 2021. Lobster. https://github.com/lobsters/lobsters.

[47] Michael Stillger, Guy M. Lohman, Volker Markl, and Mokhtar Kandil. 2001. LEO - DB2's LEarning Optimizer. In *Proceedings of the 27th International Conference on Very Large Data Bases (VLDB '01)*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 19–28.

[48] Guoxiang Tan. 2017. PERF: Avoid 'NOT IN (<subquery>)' which can get really slow. https://github.com/discourse/discourse/commit/28148197d6467cdc7469409f961c00d4e32f4c41.

[49] Zhaoguo Wang, Zhou Zhou, et al. 2022. WeTune: Automatic Discovery and Verification of Query Rewrite Rules (The Extended Version). https://ipads.se.sjtu.edu.cn/_media/publications/wtune_extend.pdf.

[50] Qi Zhou, Joy Arulraj, Shamkant Navathe, William Harris, and Jinpeng Wu. 2020. SPES: A Two-Stage Query Equivalence Verifier. *arXiv preprint arXiv:2004.00481* (2020).

[51] Qi Zhou, Joy Arulraj, Shamkant Navathe, William Harris, and Dong Xu. 2019. Automated verification of query equivalence using satisfiability modulo theories. *Proceedings of the VLDB Endowment* 12, 11 (2019), 1276–1288.

[52] Xuanhe Zhou, Guoliang Li, Chengliang Chai, and Jianhua Feng. 2021. A Learned Query Rewrite System Using Monte Carlo Tree Search. *Proc. VLDB Endow.* 15, 1 (sep 2021), 46–58. https://doi.org/10.14778/3485450.3485456