Using SAT and SQL for Pattern Mining in Relational Databases
Emmanuel Coquery, Jean-Marc Petit, Lakhdar Saïs

To cite this version:
Emmanuel Coquery, Jean-Marc Petit, Lakhdar Saïs. Using SAT and SQL for Pattern Mining in Relational Databases. 1st Workshop on COmbining COstraint solving with MIning and LEarning ('12), 2012, Montpellier, France. pp.41-46. hal-00870949

HAL Id: hal-00870949
https://hal.archives-ouvertes.fr/hal-00870949
Submitted on 19 Jan 2021

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
Using SAT and SQL for Pattern Mining in Relational Databases

Emmanuel Coquery\(^1\) and Jean-Marc Petit\(^1\) and Lakhdar Sais\(^2\)

Abstract. In this paper, we present an ongoing work bridging the gap between pattern mining, SQL, and SAT for a particular class of patterns. We extend the work presented in [2] that proposes a logical query language for rule patterns satisfying Armstrong’s axioms. Our contributions are the following: firstly, we allow a large part of the relational tuple calculus (SQL) to be used in the specification of queries. Secondly, we propose a boolean encoding of the query that can be used to compute answers even in the case of non Armstrong-compliant queries. Some experiments have been performed on top of Derby (embedded Java DBMS) and a modified version of MiniSat to show the feasibility of the approach.

1 INTRODUCTION

Declarative approaches for pattern mining attracted a growing attention in the recent years. On one hand, a way to increase declarativity is to devise high level query languages for data mining [10, 13, 14, 5, 8, 11, 2]. On the other hand, more declarativity often induce a greater expressivity, at the cost of a reduced efficiency. Constraint programming approaches for pattern mining, initiated in [15], were proposed to provide a good compromise between expressivity and efficiency.

In this paper, we propose to use SQL and SAT together for pattern mining. We extend the work presented in [2] that proposes a logical query language for rule patterns satisfying Armstrong’s axioms. While such patterns can be enumerated with efficient algorithms they might be seen as too restrictive. The first contribution of this paper is to allow a large part of the relational tuple calculus (SQL) to be used in the specification of queries. The second one is the use of a SAT solver for enumerating patterns even in the case of non Armstrong-compliant queries.

The rest of this paper is organized as follows. Section 2 presents the \(\mathcal{RLT}\) query language. Section 3 presents the basic principles of the query boolean encoding. Implementation principles are presented in section 4 together with a few optimizations, while section 5 provides some ways to reduce the number of answers. Finally some experimental results are presented in section 6 and we conclude in section 7.

2 THE \(\mathcal{RLT}\) LANGUAGE

In this section, we introduce the syntax and semantics of the \(\mathcal{RLT}\) language, which is based on a mixture of tuple relational calculus [1] and \(\mathcal{RL}\) language [2].

2.1 Preliminaries

We introduce the following definitions and notations used in the \(\mathcal{RLT}\) language:

- \(U\) is a set of attributes, noted \(\bar{A}, \bar{B}, \ldots\),
- \(\mathcal{CST}\) is a set of constants,
- \(\mathcal{CMP}\) is a set of binary\(^3\) comparisons over \(\mathcal{CST}\). It is assumed that if \(c\) and \(c'\) are two constants, and if \(\square \in \mathcal{CMP}\), then \(\bar{c} \square \bar{c}'\) is computable in constant time.
- a schema \(R\) is a finite, nonempty set of attributes from \(U\),
- a tuple \(t\) over a schema \(R\) is a total function from \(R\) to \(\mathcal{CST}\),
- \(\bar{t}[\bar{A}]\) denotes the value of \(\bar{t}\) for attribute \(\bar{A}\),
- a relation \(\bar{r}\) over a schema \(R\) is a set of tuples over \(R\).
- \(s, t, u, s_1, \ldots\) are tuple variables,
- \(A, B, C, A_1, B_1, \ldots\) are attribute variables, i.e. capital letters from the beginning of the alphabet,
- \(X, Y, Z, X_1, Y_1, \ldots\) are schema variables, i.e. capital letters from the end of the alphabet,
- \(r, r_1, r', \ldots\) are relation symbols.

To avoid ambiguity with variables, we shall use the following notations for attributes, set of attributes and tuples:

- \(\bar{A}, \bar{B}, \bar{C}, A_1, B_1, \ldots\) are single attributes,
- \(X, Y, Z, X_1, Y_1, \ldots\) are set of attributes,
- \(s, t, u, s_1, t_1, t_2, \ldots\) are tuples,
- \(\bar{c}, \bar{c}_1, \bar{c}'\) are constants.

For any function \(f : E \rightarrow E'\), we denote by \(f[e := e']\), where \(e \in E\) and \(e' \in E'\), the function \(f'\) which maps \(e\) to \(e'\) and maps \(e_1\) to \(f(e_1)\) if \(e \neq e_1\).

For any function \(f : E \rightarrow E'\), and any subset \(E_1 \subseteq E\), we denote by \(f|_{E_1}\) the restriction \(f' : E_1 \rightarrow E'\) of \(f\) to \(E_1\), which maps \(e_1\) to \(f(e_1)\).

2.2 \(\mathcal{RL}\) Formulas

This section recalls the syntax and semantics of the \(\mathcal{RL}\) language [2]. In this paper, we restrict the use of tuple variable quantifiers, by removing them from definition 1 and re-introducing them in section 2.4. We also limit the use of attribute quantifiers to some form of restricted quantification.

---

\(^1\) Université de Lyon, CNRS, France, email: first-name.lastname@liris.cnrs.fr

\(^2\) Université d’Artois, CNRS, France, email: sais@cril.univ-artois.fr

\(^3\) the binary restriction can be easily lifted, allowing arbitrary SQL boolean expressions
Let $A, B$ be attribute variables, $t, s$ tuple variables, $c$ a constant, $X$ a schema variable.

**Definition 1** The set of $\mathcal{RL}$-formulas, noted $\delta, \delta_1, \delta_2, \ldots$, is inductively defined as the smallest set verifying:

- $t.A \Box \Box c$, $t.A \Box \Box s.B$, $A = B$ and $A = \bar{A}$ are atomic $\mathcal{RL}$-formulas.
- If $\delta$ is a $\mathcal{RL}$-formula and $A$ an attribute variable, $\forall A(X)(\delta)$ is a $\mathcal{RL}$-formula.
- If $\delta$ is a $\mathcal{RL}$-formula and $A$ an attribute variable, $\exists A(X)(\delta)$ is a $\mathcal{RL}$-formula.
- If $\delta_1$ and $\delta_2$ are $\mathcal{RL}$-formulas, then $\neg \delta_1$ and $(\delta_1 \lor \delta_2)$ are $\mathcal{RL}$-formulas.

Other logical connectors such as $\lor, \Rightarrow$ and abbreviations true, false are defined as usual.

**Definition 2** A $\mathcal{RL}$-interpretation is a quadruplet $(R, \Sigma, \sigma, \tau)$ where:

- $R \subseteq U$ is a schema,
- $\Sigma$, the schema interpretation, is a function mapping each schema variable $X$ to a subset of $R$,
- $\sigma$, the attribute interpretation, is a function mapping each attribute variable $A$ to an attribute $\bar{A} \in R$,
- $\tau$, the tuple interpretation, is a function mapping each tuple variable $t$ to a tuple $\bar{t}$ over $R$.

**Definition 3** Let $\delta$ be a $\mathcal{RL}$-formula. The satisfaction of $\delta$ with respect to a $\mathcal{RL}$ interpretation $(R, \Sigma, \sigma, \tau)$, denoted by $(R, \Sigma, \sigma, \tau) \models \delta$, is defined inductively as follows:

- $(R, \Sigma, \sigma, \tau) \models t.A \Box \Box c$ if $\tau(t)[\sigma(A)] \Box \Box c$.
- $(R, \Sigma, \sigma, \tau) \models t.A \Box \Box s.B$ if $\tau(t)[\sigma(A)] \Box \Box \tau(t)[\sigma(B)]$.
- $(R, \Sigma, \sigma, \tau) \models A = \bar{A}$ if $\sigma(A) = \bar{A}$.
- $(R, \Sigma, \sigma, \tau) \models A = B$ if $\sigma(A) = \sigma(B)$.
- $(R, \Sigma, \sigma, \tau) \models \forall A(X)(\delta)$ if for all $\bar{A} \in \Sigma(X)$, $(R, \Sigma, \sigma[A := \bar{A}], \tau) \models \delta$.
- $(R, \Sigma, \sigma, \tau) \models \exists A(X)(\delta)$ if for some $\bar{A} \in \Sigma(X)$, $(R, \Sigma, \sigma[A := \bar{A}], \tau) \models \delta$.
- $(R, \Sigma, \sigma, \tau) \models \neg \delta$ if $(R, \Sigma, \sigma, \tau) \not\models \delta$.
- $(R, \Sigma, \sigma, \tau) \models (\delta_1 \land \delta_2)$ if $(R, \Sigma, \sigma, \tau) \models \delta_1$ and $(R, \Sigma, \sigma, \tau) \models \delta_2$.

**2.3 Relational Calculus**

Here we recall some definitions of the tuple relational calculus [1] (abbreviated TRC in the following). Let $\bar{A}, \bar{B}$ be attributes, $t, s$ be tuple variables, $c$ be a constant and $r$ be a relation symbol.

**Definition 4** The set of TRC-formulas, noted $\psi, \psi_1, \psi_2, \ldots$, is inductively defined as the smallest set verifying:

- $t.\bar{A} \Box \Box c$ and $t.\bar{A} \Box \Box \bar{B}$ are atomic TRC-formula.
- $r(t)$ is an atomic TRC-formula.
- If $\psi$ is a TRC-formula, and $t$ is a tuple variable then $\exists t(\psi)$ is a TRC-formula.
- If $\psi_1$ and $\psi_2$ are TRC-formulas then $(\psi_1) \land (\psi_2)$ and $(\neg \psi_1)$ are TRC formulas.

Other logical connectors such as $\lor, \Rightarrow, \forall$ and abbreviations true, false are defined as usual.

Now we recall the logical semantics of TRC-formulas. For sake of simplicity, we assume that all relations and schemas are defined over the same schema $R$. This restriction can easily be lifted though.

**Definition 5** A TRC interpretation is a pair $(d, \tau)$ where:

- $d$ is a function, the database, mapping each relation symbol $r$ to a relation $\bar{r}$ over $R$.
- $\tau$ is a function, the tuple interpretation, mapping each tuple variable $t$ to a tuple $\bar{t}$ over $R$.

**Definition 6** Let $\psi$ be a TRC-formula. The satisfaction of $\psi$ with respect to a TRC interpretation $(d, \tau)$, denoted $(d, \tau) \models \psi$, is inductively defined as follows:

- $(d, \tau) \models t.\bar{A} \Box \Box c$ if $\tau(t)[\bar{A}] \Box \Box c$.
- $(d, \tau) \models t.\bar{A} \Box \Box t.\bar{B}$ if $\tau(t)[\bar{A}] = \tau(t)[\bar{B}]$.
- $(d, \tau) \models r(t)$ if $\tau(t) \in d(r)$.
- $(d, \tau) \models \neg (\psi)$ if $(d, \tau) \not\models \psi$.
- $(d, \tau) \models \exists \bar{v}(\psi)$ if there exists a tuple $\bar{v}$ over $R$ such that $(d, \tau)[\bar{v} := \bar{v}] \models \psi$.
- $(d, \tau) \models (\psi_1) \land (\psi_2)$ if $(d, \tau) \models \psi_1$ and $(d, \tau) \models \psi_2$.

In the rest of the paper, we restrict ourselves to authorized relational calculus [1], a syntactical restriction of the relation calculus which guarantees domain independence. That is, given a database $d$, the set of tuple interpretations $\tau$ such that $(d, \tau) \models \psi$ only depends on $d$ and $\psi$.

**Definition 7** Given a TRC formula $\psi$, with $t_1, \ldots, t_k$ as free variables, the answer of $\psi$ w.r.t. a database $d$, denoted by ans$(\psi, d)$, is defined as:

$$\{ \bar{t}(t_1, \ldots, t_k) \mid (d, \tau) \models \psi \}$$

Note that for an authorized TRC formula $\psi$ and a database $d$ that associate only finite relations to relation symbols, the answer ans$(\psi, d)$ is finite.

**2.4 $\mathcal{RLT}$ Queries**

We introduce $\mathcal{RLT}$ queries, which constitute the $\mathcal{RLT}$-language.

**Definition 8** A $\mathcal{RLT}$ query is of the form:

$$\{ X_1, \ldots, X_n : R \mid \forall t_1 \ldots \forall t_k \psi \Rightarrow \delta \}$$

where:

- $X_1, \ldots, X_n$ are schema variables.
- $t_1, \ldots, t_k$ are tuple variables over the same schema $R$.
- $\psi$ is an authorized TRC-formula, that has exactly $t_1, \ldots, t_k$ as free variables.
- $\delta$ is a $\mathcal{RL}$-formula in which:
  - the only (free) tuple variables are $t_1, \ldots, t_k$.
  - the only (free) schema variables are $X_1, \ldots, X_n$.
  - there is no free attribute variable.

$\forall t_1 \ldots \forall t_k \psi \Rightarrow \delta$ is said to be the ($\mathcal{RLT}$) formula of the query.
In order to give an idea of how querying can be done using \(\mathcal{RLT}\), let us consider the following query, which finds functional dependencies in a relation \(r\) over \(R\):

\[
Q_1 = \{ (X,Y) : R \mid \forall t \forall s \forall r(s) \Rightarrow (\forall A(X)(t.A = s.A) \Rightarrow (\forall B(Y)(t.B = s.B)) \}
\]

Non Armstrong-compliant queries can also be expressed, such as:

\[
Q_2 = \{ (X,Y) : R \mid \forall t \forall s \forall r(s) \Rightarrow (\exists A(X)(t.A = s.A) \Rightarrow (\exists B(Y)(t.B = s.B)) \}
\]

or the following query for finding influence of genes:

\[
\forall \forall \cdots \forall \exists \exists \algorithm*{\text{assuming that}}
\]

In order to find answers of a \(Q\) w.r.t. \(\Sigma\), a tuple interpretation is considered to be interesting of the query and the database into a boolean formula. The domain, that is the boolean variables encoding an answer in \(\text{ans}(Q,d)\), is defined straightforwardly as follows: for each schema variable \(X \in \{X_1, \ldots, X_n\}\) and each attribute \(A \in R\), the boolean variable \(p^X_A\) is true whenever \(A \in \Sigma(X)\).

The following definition explains how the boolean formula is built from the \(\mathcal{RL}\) formula. This definition relies on an attribute interpretation. However, this interpretation has no influence on top-level \(\mathcal{RL}\) formulas, as they have no free attribute variable.

**Definition 12** Given a \(\mathcal{RL}\) formula \(\delta\), an attribute interpretation \(\sigma\) and a set of attributes \(R\), the boolean encoding of \(\delta\), denoted by \(\text{enc}(\delta, \tau, \sigma, R)\), is inductively defined as:

- \(\text{enc}(\text{true}, \tau, \sigma, R) = \text{true}\) if \(\tau = \text{true}\), \(\text{false}\) otherwise
- \(\text{enc}(\text{false}, \tau, \sigma, R) = \text{false}\) if \(\tau = \text{false}\), \(\text{true}\) otherwise
- \(\text{enc}(\tau_1 \land \tau_2, \tau, \sigma, R) = \text{true}\) if \(\tau = \text{true}\), \(\text{false}\) otherwise
- \(\text{enc}(\bar{\tau}, \tau, \sigma, R) = \text{false}\) if \(\tau = \text{true}\), \(\text{true}\) otherwise
- \(\text{enc}(\exists \Sigma(X)\delta, \tau, \sigma, R) = \bigwedge_{A \in R} p^X_A \Rightarrow \text{false}\) if \(\exists \Sigma(X)\delta = \text{false}\)

**Definition 13** A boolean interpretation \(I\) is a function from boolean variables to \(\{\text{true}, \text{false}\}\). It satisfies a boolean formula \(\gamma\), denoted by \(I \models \gamma\), if the formula obtained by replacing each variable \(p\) by \(I(p)\) is equivalent to \(\text{true}\) in the boolean algebra.

**Property 1** Let \(\Sigma\) be a schema interpretation, \(\delta\) a \(\mathcal{RL}\) formula, \(\tau\) a tuple interpretation, \(\sigma\) an attribute interpretation and \(R\) a set of attributes. Let \(I_\Sigma\) be a boolean interpretation such that for any schema variable \(X\) and any attribute \(A\), \(I_\Sigma(p^X_A) = \text{true}\) if and only if \(A \in \Sigma(X)\).

Then \((R, \Sigma, \sigma, \tau) \models \delta\) if and only if \(I_\Sigma \models \text{enc}(\delta, \tau, \sigma, R)\).

**Proof** By definitions 3 and 12.

**Definition 14** The boolean encoding of a query \(Q = \{ (X_1, \ldots, X_n) : R \mid \forall t \cdots \forall k \psi \Rightarrow \delta \}\) w.r.t. a database \(d\), denoted by \(\text{enc}(Q,d)\), is defined as:

\[
\bigwedge_{\tau \in \text{ans}(\psi,d)} \text{enc}(\delta, \tau, \sigma, R)
\]

where \(\sigma\) is any attribute interpretation\(^4\).  

**Property 2** Let \(Q = \{ (X_1, \ldots, X_n) : R \mid \forall t \cdots \forall k \psi \Rightarrow \delta \}\) be a \(\mathcal{RL}\) query, \(d\) a database and \(\Sigma\) a schema interpretation. Let \(I_\Sigma\) be a boolean interpretation such that for any schema variable \(X\) and any attribute \(A\), \(I_\Sigma(p^X_A) = \text{true}\) if and only if \(A \in \Sigma(X)\).

\(\langle \Sigma(X_1), \ldots, \Sigma(X_n) \rangle \in \text{ans}(Q,d)\) if and only if \(I_\Sigma \models \text{enc}(Q,d)\).

\(^4\) \(\sigma\) is not actually used in the encoding since \(\delta\) is closed w.r.t. attribute variables.
The cost of evaluating a RLT query using a boolean formula can be evaluated by the size of the formula and its number of boolean variables. Except for quantifiers, each construction of RLT formula generate only a constant amount of additional symbols in the encoding. For each use of a quantifier \( \Box A(X)_\delta \) if the size of \( enc(\delta, \tau, \sigma, R) \) is \( n \), then the size of \( enc(\Box A(X)_\delta, \tau, \sigma, R) \) is \( O(|R| \times n) \).

Let us consider a RLT query \( Q = \{ (X_1, \ldots, X_n) : R | \forall t_1 \ldots \forall t_k \ \psi \Rightarrow \delta \} \). Let \( n \in \mathbb{N} \) be the maximal number of quantifiers on a branch of the abstract syntax tree of \( \delta \). Then, an upper bound on the size of \( enc(Q, d) \) is \( O(|\text{ans}(\psi, d)| \times |\delta| \times |R|^{n+3}) \).

Let \( time_{\psi, d} \) be the time required to evaluate \( \text{ans}(\psi, d) \). Then an upper bound on the time complexity of the evaluation of a RLT query is \( time_{\psi, d} + O(|\text{ans}(\psi, d)| \times |\delta| \times |R|^{n+3}) \).

4 IMPLEMENTATION AND OPTIMIZATION

In this section we present the principles used in the implementation of RLT, as well as a few optimizations that help to drastically reduce the size of generated formulas. In this section, we assume given both RLT query \( Q = \{ (X_1, \ldots, X_n) : R | \forall t_1 \ldots \forall t_k \ \psi \Rightarrow \delta \} \) and database \( d \). Figure 1 presents the different steps used in the computation of answers.

4.1 Naïve translation

The first step in generating \( enc(Q, d) \) is to evaluate answers \( \text{ans}(\psi, d) \). Since \( \psi \) is an authorized TRC formula, it can be evaluated using a SQL engine, at the cost of an attribute renaming to avoid attribute name clashes between tuples. We will see in section 4.3 that in the final implementation this problem disappears. Using SQL allows to easily extend the comparison predicates and expressions used in the TRC formula \( \psi \). Then for each tuple combination \( \tau \) we generate \( enc(\delta, \tau, \sigma, R) \), with \( \sigma \) being uninitialized\(^5\).

4.2 Getting Answers From Boolean Formula

Because of property 2, the answers \( \text{ans}(Q, d) \) can be obtained by the boolean interpretations satisfying the formula \( enc(Q, d) \). We use a modified SAT-solver [6] based on MiniSat [7]. Using a SAT solver requires the formula to be translated into conjunctive normal form (CNF). For this we use a linear translation based on [16]. This translation propagates constants through standard logical equivalences such as \( \eta_1 \land \text{true} \equiv \eta_1 \) and \( \eta_1 \land \text{false} \equiv \text{false} \). The translation into CNF also introduce new variables. However, the value of these new variables can be deduced from the values of the variables in the original formula. This information can be transmitted to the enumerating SAT solver, allowing it to branch only on variables of the original formula. Moreover the use of modern SAT solvers allows to benefit from efficient propagation techniques, learning [12, 17] and dynamic search heuristics [17, 4]. For an extensive overview of current techniques to solve SAT, the reader is referred to [3].

4.3 Caching Subformulas

By looking at definition 12, one can see that the generated formula for each tuple combination depend on the structure of \( \delta \) and on the evaluation of the atomic sub formulas of \( \delta \) for (all) the possible attribute interpretations \( \sigma \). That is, the actual value of tuples is not important, only the value they give to atomic formulas in \( \delta \) matters.

For each atomic formula \( \delta_i \) in \( \delta \), we introduce a series of boolean variables \( g_{\delta_i}^\delta \) for all possible attribute interpretation \( \sigma : A \rightarrow R \) where \( A \) is the set of attribute variables appearing in \( \delta_i \).

Definition 15 The intermediate encoding \( enc'(\delta, \sigma, R) \) is inductively defined as follows:

- \( enc'(t.A = c, \sigma, R) = q_{t.A}^{\delta} \)
- \( enc'(t.A = s, \sigma, R) = q_{t.A}^{\delta} \)
- \( enc'(A = \bar{A}, \sigma, R) = \text{true} \) if \( \sigma(A) = \bar{A} \)
- \( enc'(A = B, \sigma, R) = \text{true} \) if \( \sigma(A) = \sigma(B) \), false otherwise
- \( enc'(-\delta, \sigma, R) = -enc'(\delta, \sigma, R) \)
- \( enc'(\delta_1 \land \delta_2, \sigma, R) = enc'(\delta_1, \sigma, R) \land enc'(\delta_2, \sigma, R) \)
- \( enc'(\forall A(X)_\delta, \sigma, R) = \bigwedge_{\bar{A} \in R} \left( p_{\bar{A}}^\delta \Rightarrow enc'(\delta, \sigma[A := \bar{A}], R) \right) \)
- \( enc'(\exists A(X)_\delta, \sigma, R) = \bigvee_{\bar{A} \in R} \left( p_{\bar{A}}^\delta \land enc'(\delta, \sigma[A := \bar{A}], R) \right) \)

This definition is similar to definition 12, except for atomic formulas where tuple variables appear. By construction, for a given tuple interpretation \( \tau \), \( enc(\delta, \tau, \sigma, R) \) can be obtained by replacing each variable \( q_{\delta_i}^{\delta} \) in \( enc'(\delta, \sigma, R) \) by \( enc(\delta_i, \tau, \sigma, R) \).

This suggests a change in the generation of \( enc(Q, d) \): the SQL engine can be used to evaluate the value of atomic formulas in \( \delta \) w.r.t. all relevant attribute interpretations \( \sigma \), that is w.r.t. all combination of values for attribute variables appearing in the sub formula. The encoding of the RLT formula for this tuple combination is then obtained by propagating these boolean values in the intermediate encoding.

The benefits of this change are twofold. Firstly, the comparison operator and expressions used in atomic RLT formulas can easily be extended to any operator supported by the SQL engine. Secondly, and more importantly, given two tuple interpretations \( \tau_1 \) and \( \tau_2 \), if the evaluation of all atomic formulas w.r.t. all attribute interpretation are the same for \( \tau_1 \) and \( \tau_2 \), then \( enc(\delta, \tau_1, \sigma, R) = enc(\delta, \tau_2, \sigma, R) \). Since the occurrence of

\(^5\) In fact, we just initialize the data structure for representing the function.
these two formulas are used in same conjunction, one of them is useless and can be discarded. This discarding behavior can be obtained, either by the use of the DISTINCT keyword in the SQL query, or more efficiently, by the use of a prefix tree to store and search for atomic formula evaluations. This optimization can be essential for the diminution of the size of the generated formula as shown in section 6.

4.4 Attribute Variable Combinatorics

Another way to reduce the size of enc(Q, d) is to try to reduce the number of nested attribute quantifiers in δ. We assume, without loss of generality, that each attribute variable appears exactly in one quantifier in δ. The attribute quantifiers can be “pushed down” towards atomic formulas in δ, by using the following standard logical equivalences:

- ∀A(X)δ1 ∧ δ2 ≡ (∀A(X)δ1) ∧ δ2 if A does not appear in δ2
- ∃A(X)δ1 ≡ ¬∀A(X)δ1
- ∀A(X)¬B(Y)δ1 ≡ ¬∀B(Y)∀A(X)δ1

For example, using these equivalences ∃A(X)¬B(Y)(t. A = 1 ⇒ t.B = 1) ≡ (∃A(X)t. A = 1) ⇒ (∃B(Y).t = B = 1). The size of the generated formula in the second case is O(|R|) smaller than the one generated in the first case.

The use of ∧ commutativity and associativity may allow for more optimizations such as ∀A(X)¬B(Y)(δ1 ∧ (δ2 ∧ δ3)) ≡ ∀B(Y)(δ2) ∧ ∀A(X)(δ1 ∧ δ3) if A does not appear in δ2 and B does not appear in δ1 nor δ3. From this point of view, this kind of optimization can be brought near rule based optimization in relational queries [1].

5 REDUCING RESULT SIZE

As the search space size is $2^n \times |R|$, it is interesting to reduce the number of results. For example, it is usual when mining functional dependencies to output a minimal base of rules from which all rules can be inferred using Armstrong’s axioms [9]. However, since our language is not supposed to be Armstrong-compliant [2], we express a wider class of queries without knowing a priori whether or not a given property is true (e.g. transitivity or reflexivity). Thus a canonical condensed representation of rules may not exist. Nevertheless, we provide means to end-users to reduce the number of results, while keeping interesting information. These means come in two flavors: firstly constraining the resulting sets of attributes, and secondly output only minimal sets (or maximal) w.r.t. set inclusion for some schema variables.

5.1 Constraining schema variables

The following examples illustrate how $\mathcal{RLC}$ formulas can be used to constrain schema variables. Assume one wants to constraint two schema variables $X$ and $Y$, such that $\Sigma(X) \cap \Sigma(Y) = \emptyset$. This constraint can be expressed by $\forall A(X) \forall B(Y) \neg A = B$. The formula $\exists A(X)$ true imposes that $\Sigma(X)$ contains at least one attribute, while the formula $\forall A(X) \forall B(Y) A = B$ imposes that $\Sigma(X)$ contains at most one attribute.

One can remark that $(R, \Sigma, \sigma, \tau) \models \exists A(X) X = \bar{A}$ if and only if $\bar{A} \in \Sigma(X)$. Therefore $\text{enc}(\exists A(X) X = \bar{A}, \sigma, \tau, R) \equiv p_{\bar{A}}^{X^5}$. This allows constraining schema variables by using any boolean formula on the variables $p_{\bar{A}}^{XY}$. The following examples illustrate how $\mathcal{RLC}$ formulas can be used to constrain schema variables. Assume one wants to constraint two schema variables $X$ and $Y$, such that $\Sigma(X) \cap \Sigma(Y) = \emptyset$. This constraint can be expressed by $\forall A(X) \forall B(Y) \neg A = B$. The formula $\exists A(X)$ true imposes that $\Sigma(X)$ contains at least one attribute, while the formula $\forall A(X) \forall B(Y) A = B$ imposes that $\Sigma(X)$ contains at most one attribute.

One can remark that $(R, \Sigma, \sigma, \tau) \models \exists A(X) X = \bar{A}$ if and only if $\bar{A} \in \Sigma(X)$. Therefore $\text{enc}(\exists A(X) X = \bar{A}, \sigma, \tau, R) \equiv p_{\bar{A}}^{X^5}$. This allows constraining schema variables by using any boolean formula on the variables $p_{\bar{A}}^{XY}$. The following examples illustrate how $\mathcal{RLC}$ formulas can be used to constrain schema variables. Assume one wants to constraint two schema variables $X$ and $Y$, such that $\Sigma(X) \cap \Sigma(Y) = \emptyset$. This constraint can be expressed by $\forall A(X) \forall B(Y) \neg A = B$. The formula $\exists A(X)$ true imposes that $\Sigma(X)$ contains at least one attribute, while the formula $\forall A(X) \forall B(Y) A = B$ imposes that $\Sigma(X)$ contains at most one attribute.

One can remark that $(R, \Sigma, \sigma, \tau) \models \exists A(X) X = \bar{A}$ if and only if $\bar{A} \in \Sigma(X)$. Therefore $\text{enc}(\exists A(X) X = \bar{A}, \sigma, \tau, R) \equiv p_{\bar{A}}^{X^5}$. This allows constraining schema variables by using any boolean formula on the variables $p_{\bar{A}}^{XY}$ through the $\mathcal{RLC}$ formula of an $\mathcal{RLT}$ query. A consequence of this remark is that given $\mathcal{RLC}$ formula $\exists A(X)$ true imposes that $\Sigma(X)$ contains at least one attribute, while the formula $\forall A(X) \forall B(Y) A = B$ imposes that $\Sigma(X)$ contains at most one attribute.

One can remark that $(R, \Sigma, \sigma, \tau) \models \exists A(X) X = \bar{A}$ if and only if $\bar{A} \in \Sigma(X)$. Therefore $\text{enc}(\exists A(X) X = \bar{A}, \sigma, \tau, R) \equiv p_{\bar{A}}^{X^5}$. This allows constraining schema variables by using any boolean formula on the variables $p_{\bar{A}}^{XY}$ through the $\mathcal{RLC}$ formula of an $\mathcal{RLT}$ query. A consequence of this remark is that given $\mathcal{RLC}$ formula $\exists A(X)$ true imposes that $\Sigma(X)$ contains at least one attribute, while the formula $\forall A(X) \forall B(Y) A = B$ imposes that $\Sigma(X)$ contains at most one attribute.

One can remark that $(R, \Sigma, \sigma, \tau) \models \exists A(X) X = \bar{A}$ if and only if $\bar{A} \in \Sigma(X)$. Therefore $\text{enc}(\exists A(X) X = \bar{A}, \sigma, \tau, R) \equiv p_{\bar{A}}^{X^5}$. This allows constraining schema variables by using any boolean formula on the variables $p_{\bar{A}}^{XY}$ through the $\mathcal{RLC}$ formula of an $\mathcal{RLT}$ query. A consequence of this remark is that given $\mathcal{RLC}$ formula $\exists A(X)$ true imposes that $\Sigma(X)$ contains at least one attribute, while the formula $\forall A(X) \forall B(Y) A = B$ imposes that $\Sigma(X)$ contains at most one attribute.

One can remark that $(R, \Sigma, \sigma, \tau) \models \exists A(X) X = \bar{A}$ if and only if $\bar{A} \in \Sigma(X)$. Therefore $\text{enc}(\exists A(X) X = \bar{A}, \sigma, \tau, R) \equiv p_{\bar{A}}^{X^5}$. This allows constraining schema variables by using any boolean formula on the variables $p_{\bar{A}}^{XY}$ through the $\mathcal{RLC}$ formula of an $\mathcal{RLT}$ query. A consequence of this remark is that given $\mathcal{RLC}$ formula $\exists A(X)$ true imposes that $\Sigma(X)$ contains at least one attribute, while the formula $\forall A(X) \forall B(Y) A = B$ imposes that $\Sigma(X)$ contains at most one attribute.
of attributes in $R$, i.e. only $m$ attributes were kept in the relation $r$. As expected, the CPU time increases exponentially w.r.t. the number of attributes, as it increases both the search space and the size of the boolean formula. Figure 3 shows the size of the generated CNF (in number of variable occurrences) and the number of attribute interpretations w.r.t. the number of tuples, the query being run on the 27 attributes of the relation. As expected, the number of interpretations grows quadratically as there are two uncorrelated tuple variables in the query. One can remark that the size of the CNF, increases slowly. This is due to the low number of useful additional tuple interpretations. Indeed the size of the generated CNF increases proportionally to the number of useful tuple interpretations. This shows the efficiency of the cache optimization on boolean formula generation.

7 CONCLUSION

We presented an ongoing work on the query language $RLT$ for pattern mining. Namely, we presented the semantics of the language, as well as a translation of queries and data into a boolean formula. Implementation techniques used for implementing a query engine were presented and some experimental results show the feasibility of the approach.

Several issues remain to be explored. One the theoretical side, it would be interesting to characterize the complexity of answering $RLT$ queries (e.g. the complexity of determining the emptiness of a query). One the practical side, performances of the query engine and query optimization techniques have to be investigated through a comprehensive set of databases. The performance of the current implementation could be improved either through high level optimization in order to generate better, more easy to solve, formulas, or through the elaboration of (SAT) engines dedicated to the enumeration of models of boolean formulas. An other direction of improvement is to enrich the language, for example with counting statements to be able to take into account the well-known frequency constraint in data mining. This could be treated using pseudo-boolean constraints such as in [15].

REFERENCES