Using SAT and SQL for Pattern Mining in Relational Databases

Emmanuel Coquery\textsuperscript{1} and Jean-Marc Petit\textsuperscript{1} and Lakhdar Sais\textsuperscript{2}

\section*{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 \cite{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.

\section{INTRODUCTION}

Declarative approaches for pattern mining attracted a growing attention in the recent years. On the one hand, a way to increase declarativity is to devise high level query languages for data mining \cite{10, 13, 14, 5, 8, 11, 12}. 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 \cite{15}, were proposed to provide a good compromise between expressivity and efficiency.

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 \cite{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.

The rest of this paper is organized as follows. Section 2 presents the \textit{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.

\textsuperscript{1}Université de Lyon, CNRS, France, email: first-name.lastname@liris.cnrs.fr
\textsuperscript{2}Université d’Artois, CNRS, France, email: sais@cril.univ-artois.fr

\section{THE \textit{RLT} LANGUAGE}

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

\subsection{Preliminaries}

We introduce the following definitions and notations used in the \textit{RLT} language:

- $\mathcal{U}$ is a set of attributes, noted $\bar{A}, \bar{B}, \ldots$,
- $\mathcal{CST}$ is a set of constants,
- a schema $R$ is a finite, nonempty set of attributes from $\mathcal{U}$,
- a tuple $\bar{t}$ 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:

- $\bar{A}, \bar{B}, \bar{C}, \bar{A}_1, \bar{B}_1 \ldots$ are single attributes,
- $\bar{X}, \bar{Y}, \bar{Z}, \bar{X}_1, \bar{Y}_1 \ldots$ are set of attributes,
- $\bar{s}, \bar{t}, \bar{t}_1, \bar{t}_2 \ldots$ are tuples,
- $\bar{c}, \bar{c}_1, \bar{c}'$ are constants.

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

\subsection{\textit{RL} Formulas}

This section recalls the syntax and semantics of the \textit{RL} language \cite{2}. In this paper, we restrict the used of tuple variable quantifiers, by removing them from definition 1 and reintroducing them in section 2.4. We also limit the use of attribute quantifiers to some form of restricted quantification.

Let $A, B$ be attribute variables, $t, s$ tuple variables, $\bar{c}$ a constant, $X$ a schema variable.
Definition 1: $RL$-formulas, noted $\delta, \delta_1, \delta_2, \ldots$ are inductively defined as follows:

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

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

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

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

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

- $(R, \Sigma, \sigma, \tau) \models t.A = \bar{c}$ if $\tau(t)[\sigma(A)] = \bar{c}$
- $(R, \Sigma, \sigma, \tau) \models t.A = s.B$ if $\tau(t)[\sigma(A)] = \tau(t)[\sigma(B)]$
- $(R, \Sigma, \sigma, \tau) \models A = \bar{A}$ 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, $\bar{c}$ be a constant and $r$ be a relation symbol.

Definition 4: TRC-formulas, noted $\psi, \psi_1, \psi_2, \ldots$ are inductively defined as follows:

- $t.A = \bar{c}$ and $t.A = t.B$ 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 $\lor$, $\Rightarrow$, quantifier $\forall$ 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 $t$ 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:

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

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

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 $RLT$ Queries

We introduce $RLT$ queries, which constitute the $RLT$-language.

Definition 8: A $RLT$ query is of the form:

$$\{X_1, \ldots, X_n : R \mid \forall v_1 \ldots \forall v_k \psi \Rightarrow \delta\}$$

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 an $RLT$-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.

$$\forall v_1 \ldots \forall v_k \psi \Rightarrow \delta$$

is said to be the ($RLT$) formula of the query.

In order to give an idea of how querying can be done using $RLT$, let us consider the following query, which find functional dependencies in a relation $r$ over $R$:

$$Q_1 = \{(X, Y) : R \mid \forall v_1 s/r(\psi(t) \land r(s)) \Rightarrow (\forall A(X)(t.A = s.A)) \Rightarrow (\forall B(Y)(t.B = s.B))\}$$
Non Armstrong-compliant queries can also be expressed, such as:

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

or

\[ Q_3 = \{ (X,Y) : R \mid \forall t \forall r(t) \Rightarrow (\exists A(X)(t.A = 1)) \Rightarrow (\forall B(Y)(t.B = 0)) \} \]

We now define the semantics of the language, first by defining the satisfaction of \( \mathcal{RLT} \) query formulas, and then by defining the answers of a \( \mathcal{RLT} \) query w.r.t. a given database.

**Definition 9** A \( \mathcal{RLT} \) interpretation is a triple \((d, R, \Sigma)\) where:

- \( R \subseteq \mathcal{U} \) is a set of attributes;
- \( d \) is a function, the database, mapping each relation symbol \( r \) to a relation \( \bar{r} \) over \( R \);
- \( \Sigma \), the schema interpretation, is a function mapping each schema variable \( X \) to a subset of \( R \).

**Definition 10** Let \( \zeta = \forall t_1 \ldots \forall t_k \psi \Rightarrow \delta \) be a \( \mathcal{RLT} \)-formula. \( \zeta \) satisfies a \( \mathcal{RLT} \) interpretation \((d, R, \Sigma)\), denoted \((d, R, \Sigma) \models \zeta \), if, for any tuple interpretation \( \tau \), and any attribute interpretation \( \sigma \), if \((d, \tau) \models \psi \) then \((R, \Sigma, \sigma, \tau) \models \psi \).

Taking any tuple interpretation corresponds to the use of \( \forall \) quantifiers in \( \mathcal{RLT} \) formulas. On the other hand, attribute interpretations are not important since \( \delta \) does not contain free variables.

**Definition 11** Given a database \( d \) and a \( \mathcal{RLT} \) query \( Q = \{ (X_1, \ldots , X_n) : R \mid \forall t_1 \ldots \forall t_k \psi \Rightarrow \delta \} \), the answer of \( Q \) in \( d \), denoted \( \text{ans}(Q, d) \), is defined as:

\[ \text{ans}(Q, d) = \{ \Sigma(X_1), \ldots , \Sigma(X_n) \} \mid (d, R, \Sigma) \models \psi \Rightarrow \delta \] is a \( \mathcal{RLT} \) interpretation and \((d, R, \Sigma) \models \psi \Rightarrow \delta \)

## 3 Boolean Encoding

In order to find answers of a query \( Q = \{ (X_1, \ldots , X_n) : R \mid \forall t_1 \ldots \forall t_k \psi \Rightarrow \delta \} \) w.r.t. a database \( d \), we propose an encoding of the query into the database and into the boolean algebra. More precisely, for each interesting tuple interpretation \( \tau \), we generate a boolean formula representing the truth value of \( \delta \) w.r.t. \( \Sigma \). A tuple interpretation is considered to be interesting if, together with \( d \), it satisfies \( \psi \). The boolean formula for computing \( \text{ans}(Q, d) \) is then the conjunction of the formulas for each interesting tuple interpretation.

### 3.1 Translation To Boolean Formula

**Domain** 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 \( \mathcal{RLT} \) formula. This definition relies on an attribute interpretation. However, this interpretation has no influence on top-level \( \mathcal{RLT} \) formulas, as they have no free attribute variable.

**Definition 12** Given a \( \mathcal{RL} \) formula \( \delta \), a tuple interpretation \( \tau \), 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:

- \( \text{enc}(t.A = \tilde{c}, \tau, \sigma, R) = \text{true} \), if \( \tau(\tilde{c})[\sigma(A)] = \tilde{c} \), \( \text{false} \) otherwise
- \( \text{enc}(t.A = s.B, \tau, \sigma, R) = \text{true} \), if \( \tau(s)[\sigma(B)] \), \( \text{false} \) otherwise
- \( \text{enc}(A = \tilde{A}, \tau, \sigma, R) = \text{true} \), if \( \sigma(A) = \tilde{A} \)
- \( \text{enc}(A = B, \tau, \sigma, R) = \text{true} \), if \( \sigma(A) = \sigma(B) \), \( \text{false} \) otherwise
- \( \text{enc}(\neg \delta, \tau, \sigma, R) = \neg \text{enc}(\delta, \tau, \sigma, R) \)
- \( \text{enc}(\delta_1 \land \delta_2, \tau, \sigma, R) = \text{enc}(\delta_1, \tau, \sigma, R) \land \text{enc}(\delta_2, \tau, \sigma, R) \)
- \( \text{enc}(\forall A(X)\delta, \tau, \sigma, R) = \bigwedge_{A \in R} (p_X^A \Rightarrow \text{enc}(\delta(X), \tau, \sigma[A := \tilde{A}], R)) \)
- \( \text{enc}(\exists A(X)\delta, \tau, \sigma, R) = \bigvee_{A \in R} (p_X^A \land \text{enc}(\delta(X), \tau, \sigma[A := \tilde{A}], R)) \)

**Definition 13** A boolean interpretation \( I \) is a function from boolean variables to \{true, 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 \((R, \Sigma, \sigma, \tau) \models \gamma \) if and only if \( I_\Sigma \models \text{enc}(\gamma, \tau, \sigma, R) \).

**Proof** By definitions 3 and 12.

**Definition 14** The boolean encoding of a query \( Q = \{ (X_1, \ldots , X_n) : R \mid \forall t_1 \ldots \forall t_k \psi \Rightarrow \delta \} \) w.r.t. a database \( d \), denoted by \( \text{enc}(Q, d) \), is defined as:

\[ \bigwedge_{\tau \in \text{ans}(Q, d)} \text{enc}(\delta, \tau, \sigma, R) \] where \( \sigma \) is any attribute interpretation.

**Property 2** Let \( Q = \{ (X_1, \ldots , X_n) : R \mid \forall t_1 \ldots \forall t_k \psi \Rightarrow \delta \} \) be a \( \mathcal{RLT} \) 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 \( A \in \Sigma(X) \).

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

**Proof** By definitions 11, 7 and 14, by property 1 and by remarking that if \( \tau_{(t_1, \ldots , t_k)} = \tau_{(t_1, \ldots , t_k)} \), then \( \text{enc}(\delta, \tau', \sigma, R) = \text{enc}(\delta, \tau, \sigma, R) \).

### 3.2 Theoretical Complexity

The cost of evaluating a \( \mathcal{RLT} \) query using a boolean formula can be evaluated by the size of the formula and its number

\[ \delta \]

\[ \sigma \]

\[ \text{is not actually used in the encoding since } \delta \text{ is closed w.r.t. attribute variables} \]
of boolean variables. Except for quantifiers, each construction of \( \mathcal{RL} \) 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 \( \text{enc}(\delta, \tau, \sigma, R) \) is \( n \), then the size of \( \text{enc}(\Box A(X)\delta, \tau, \sigma, R) \) is \( O(|R| \times n) \).

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

Let \( \text{time}_{\psi,d} \) be the time required to evaluate \( \text{ans}(\psi, d) \). Then an upper bound on the time complexity of the evaluation of a \( \mathcal{RLT} \) query is \( \text{time}_{\psi,d} + O(|\text{ans}(\psi, d)| \times |\delta| \times |R|^{n_{\psi_\delta}} \times 2^{n_{\psi_\delta} |R|}) \).

## 4 IMPLEMENTATION AND OPTIMIZATION

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

### 4.1 Naive translation

The first step in generating \( \text{enc}(Q, d) \) 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 \( \text{enc}(\delta, \tau, \sigma, R) \), with \( \sigma \) being uninitialized.

### 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 \( \text{enc}(Q, d) \). 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 \( \eta \land \text{true} \equiv \eta \) and \( \eta \land \text{false} \equiv \text{false} \). The translation into CNF also introduce new variables. However, the value of these new variables can be deduced from the values of the variables.

4 In fact, we just initialize the data structure for representing the function.
4.4 Attribute Variable Combinatorics

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

- $\forall A(X)(\delta_1 \land \delta_2) \equiv (\forall A(X)\delta_1) \land \delta_2$ if $A$ does not appear in $\delta_2$
- $\exists A(X)(\delta_1 \land \delta_2) \equiv (\exists A(X)\delta_1) \land \delta_2$ if $A$ does not appear in $\delta_2$
- $\exists A(X)\neg \delta_1 \equiv \neg \exists A(X)\delta_1$
- $\forall A(X)\neg \delta_1 \equiv \neg \forall A(X)\delta_1$
- $\forall A(X)\forall B(Y)\delta_1 \equiv \forall B(Y)\forall A(X)\delta_1$
- $\exists A(X)\exists B(Y)\delta_1 \equiv \exists B(Y)\exists A(X)\delta_1$

For example, using these equivalences $\exists A(X)\forall B(Y)(t.A = 1 \Rightarrow B) \equiv (\forall A(X)(t.A = 1) \Rightarrow (\forall B(Y)t.B = 1)$. 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 $\land$ commutativity and associativity may allow for more optimizations such as $\forall A(X)\forall B(Y)(\delta_1 \land (\delta_2 \land \delta_3)) \equiv \forall B(Y)(\delta_2 \land \forall A(X)(\delta_1 \land \delta_3))$ if $A$ does not appear in $\delta_2$ and $B$ does not appear in $\delta_1$ or $\delta_3$. 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 $\mathcal{RL}$ formulas can be used to constrain schema variables. Assume one wants to constrain 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 $\forall A(X)\forall B(X) A = B$ imposes that $\Sigma(X)$ contains at most one attribute.

One can remark that $(R, \text{Sigma}, \sigma, \tau) \models \exists A(X)X = A$ if and only if $A \in \Sigma(X)$. Therefore $\text{enc}(\exists A(X)X = A, \sigma, \tau, R) \equiv p_A^X$. This allows constraining schema variables by using any boolean formula on the variables $p_A^X$ through the $\mathcal{RL}$ formula of an $\mathcal{RLT}$ query. A consequence of this remark is given by the following property:

**Property 3** Given a database $d$ and a $\mathcal{RLT}$ query $Q$, the problem of determining whether $\text{ans}(Q, d) \neq \emptyset$ is NP-Hard.

5.2 Minimizing/Maximizing Schema Variables

Another way to reduce the number of results is to minimize or maximize schema interpretation values for some variables.

**Definition 16** A schema interpretation $\Sigma$ is said to be minimal (resp. maximal) w.r.t. a schema variable $X$, a database $d$ and a $\mathcal{RLT}$ query $Q = \{X_1, \ldots, X_n\} : R \mid \zeta$ if there is no $RX$ such that $RX \subset \Sigma(X)$ (resp. $RX \supset \Sigma(X)$) and $(d, R, \Sigma[X := RX]) \models \zeta$.

$\Sigma$ is said to be locally minimal (resp. maximal) w.r.t. $X$, $d$ and $Q$ if there is no $RX$ such that $RX \subset \Sigma(X)$ w.r.t. $|RX| = |\Sigma(X)| - 1$ (resp. $RX \supset \Sigma(X)$ w.r.t. $|RX| = |\Sigma(X)| + 1$) and $(d, R, \Sigma[X := RX]) \models \zeta$.

$Q$ is said to be monotone (resp. antimonotone) w.r.t. $X$, $d$ and $Q$ if there is no $RX$ such that $RX \subset \Sigma(X)$ w.r.t. $|RX|$ is no (resp. minimal) w.r.t. $X$, $d$ and $Q$ then it is maximal (resp. minimal) w.r.t. $X$, $d$ and $Q$.

We propose the following boolean encoding of the locally minimal/maximal constraint on a schema variable $X$, a database $d$ and a $\mathcal{RLT}$ query $Q = \{X_1, \ldots, X_n\} : R \mid \forall t_1, \ldots, t_k \psi \Rightarrow \delta$.

Given a boolean formula $\gamma$, we denote by $\gamma[\gamma'/p]$ the formula obtained by replacing each occurrence of $p$ in $\gamma$ by $\gamma'$.

- $\text{enc}_{\text{min}}(X, d, Q) = \bigwedge_{A \in R} p_A^X \Rightarrow \neg(\text{enc}(Q, d)[\text{false}/p_A^X])$
- $\text{enc}_{\text{max}}(X, d, Q) = \bigwedge_{A \in R} \neg p_A^X \Rightarrow \neg(\text{enc}(Q, d)[\text{true}/p_A^X])$

The size of this constraint’s boolean encoding is $|R|$ times the size of the original query’s boolean encoding.

6 PRELIMINARY EXPERIMENTS

The CNF generator has been coded in Java, while the modified MiniSat solver in C++. We have used an embedded DBMS (Derby), since it allows to include the execution of SQL statements in CPU time results. The experiments where conducted on a 2GHz dual core Athlon processor with 3GB of RAM, running Linux.

This section presents a few experimental results on the following $\mathcal{RLT}$ query:

$$\{X, Y \mid R \mid \forall t_1 \forall t_2 \exists t_1.A = t_2.A \Rightarrow (\forall B(Y)t_1.B = t_2.B) \land (\forall A(X)t_1.A = t_2.A)$$

This query finds functional dependencies $X \rightarrow Y$ in $r$, $X$ and $Y$ having an empty intersection and $Y$ being a singleton. Moreover $X$ was minimized.

The relation initially contains 2013 tuples and 27 attributes. Figure 2 shows evolution of CPU time w.r.t. the number of attributes, the 2013 tuples in $r$ being used. As expected, the CPU time increases exponentially w.r.t. the number of attributes, as it increases both the search space and the size
7 CONCLUSION

We presented an ongoing work on the query language \textit{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 where presented and some experimental results assess 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 \textit{RLT} queries (e.g. the complexity of determining the emptiness of query). One the practical side, performances of the query engine and query optimization techniques have to be investigated through comprehensive 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 optimizations of the SAT engine used to enumerate answers. 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.

REFERENCES