RQL: un langage “à la SQL” pour découvrir des règles à partir des données

Brice Chardin  
LIAS, ISAE-ENSMA  
France

Emmanuel Coquery  
Université Lyon 1  
France

Jean-Marc Petit  
INSA Lyon  
France

Marie Pailloux  
Université Blaise Pascal  
France

ABSTRACT
RQL (pour Rule Query Language) est un langage de requêtes “à la SQL” qui étend et généralise les dépendances fonctionnelles à de nouvelles catégories de règles. RQL apporte aux analystes de données un outil pratique pour découvrir les implications logiques entre attributs d’une base de données. Ces implications peuvent mettre en évidence des problèmes de qualité de données ou de nouvelles corrélations inattendues entre les attributs. Le traitement de ces requêtes RQL est basé sur une technique de réécriture qui délègue un maximum de calculs au SGBD sous-jacent. Cette contribution vise à renforcer le lien entre la fouille de données et les bases de données et de faciliter l’utilisation de techniques de fouille par des analystes ou des étudiants habitués au SQL.

1. INTRODUCTION
Pattern mining can be seen as an automated part of data exploration. For instance, functional dependencies or conditional functional dependencies are definitely useful to understand the data and to identify data quality problems [5]. However, pattern mining techniques are rarely usable directly by data analysts. Most of the time, they have to perform some data pre-processing between different systems and formats. The pattern mining codes themselves often require to be compiled from some specific programming languages. All these steps are out of reach of many data analysts, rendering round-trip engineering into a nightmare. Automated rule generation can also flood the analyst with huge amounts of patterns, and make it difficult to extract useful information. Other techniques have to be provided to interact with the data and give useful feedback to the analysts.

Demo contribution To improve pattern mining usability for data exploration, we introduce a Rule Query Language (RQL) that allows SQL-aware analysts to use pattern mining techniques with an interactive, user-friendly interface. In this demonstration, we show the usability of this web interface with the point of view of a data analyst. We focus on the expressive power of RQL through various examples, showing how easy it is to devise new and surprising rules with a very simple language derived from SQL. We also introduce how the data analysts can interact with the system through RQL queries and counterexamples taken from the database. During the demo, participants will be invited to formulate their own queries on predefined databases to discover attribute relationships through generated rules and counterexamples.

Figure 1 gives a preview of the web interface for RQL, made available¹ for research and educational purposes. This interface provides an unified access to the user’s data and pattern mining techniques using declarative languages: SQL and RQL.

From previous works [1, 2, 6], RQL is compliant with Armstrong’s axioms, i.e. the language generalizes functional dependencies to a new class of dependencies based on logical implications (if ... then statements). To the best of our knowledge, this class of dependencies has not been studied before. We have proven that these dependencies can be efficiently computed from a database using a two step technique. First, a non trivial SQL subquery is generated to compute a base of the associated closure system – a base is also called a context in formal concept analysis terminology [8]. Then, a state of the art algorithm [12] is used to generate a canonical cover of rules from this base. This approach allows RQL to benefit from DBMSs’ query optimization to access the data, and keep the data where they are.

Related works Defining specific languages for pattern mining is a long standing goal [3], for example using constraint programming techniques [10]. Nevertheless, we argue that

¹http://rql.insa-lyon.fr
pattern mining languages should benefit from direct extensions of the SQL language, since data are often stored in DBMSs. Other practical approaches, as close as possible of DBMSs, have been proposed to interact more directly with DBMSs query engines [7, 14, 4].

2. THE RQL QUERY LANGUAGE

To make things concrete, let us consider the running example given in Figure 3 with the EMP table. The attribute Educlevel represents the number of years of formal education, Sal the yearly salary, Bonus the yearly bonus and Comm the yearly commission. The meaning of other attributes is straightforward.

To begin with, let us extract functional dependencies (FD) from the relation EMP. Recall that a FD \( X \rightarrow Y \) holds in \( r \) if for all tuples \( t_1, t_2 \in r \), and for all attribute \( A \in X \) such that \( t_1[A] = t_2[A] \) then for all \( A \in Y \), \( t_1[A] = t_2[A] \). With RQL, FDs are expressed in a similar way.

Example 1. \( Q_1 \) discovers FDs from EMP over a subset of attributes.

\[ Q_1: \text{FINDRULES OVER Empno, Lastname, Workdept, Job, Sex, Bonus, Mgrno} \]

\[ \text{SCOPE t1, t2 Emp} \]

\[ \text{CONDITION ON } \$A \text{ IS t1.}$A = t2.$A \]

Note how the \text{CONDITION} clause matches the previous logical implication. We have also restricted FDs discovery to a subset of seven attributes in the \text{OVER} clause. In this example, a canonical cover of FDs that hold in EMP is generated (composed of twenty-four FDs), including FDs such as \( \text{Empno} \rightarrow \text{Lastname or Workdept} \rightarrow \text{Job} \).

Overall, a RQL query has the following general form:

\[ \text{FINDRULES OVER } \{ \text{set of attributes: } A_1, \ldots, A_n \} \]

\[ \text{SCOPE } \{ \text{tuple variables: } t_1, \ldots, t_n \} \]

\[ \text{WHERE } \{ \text{condition on } (t_1, \ldots, t_n) \} \]

\[ \text{CONDITION ON } \{ \text{attribute variable: } \$A \} \]

\[ \text{IS } \{ \text{condition on } (A_1, t_1, \ldots, t_n) \} \]

The \text{FINDRULES} keyword identifies a RQL query, which generates rules of the form \( X \rightarrow Y \) with \( X \) and \( Y \) disjoint attribute sets taken from the \text{OVER} clause. The \text{SCOPE} clause defines tuple-variables over some tables obtained by classical SQL queries. An optional \text{WHERE} clause defines relationships between tuple-variables, similar to the SQL \text{WHERE} clause. The \text{CONDITION ON $A$} clause defines the predicate to be satisfied by each attribute $A$ occurring in the left- and right-hand sides of the rule.

To illustrate the expressiveness of RQL queries, we now provide several examples.

Example 2. Let us consider null values known to be common in real-life databases. With RQL, the data analyst has the opportunity to discover rules between attributes with respect to null values as shown with query \( Q_2 \).

\[ Q_2: \text{FINDRULES OVER Empno, Lastname, Workdept, Job, Sex, Bonus, Mgrno} \]

\[ \text{SCOPE t1 Emp} \]

\[ \text{CONDITION ON } \$A \text{ IS t1.$A IS NULL} \]

The rule \( Mgrno \rightarrow Workdept \) holds in EMP since each time the attribute \( Mgrno \) is null in a tuple, then \( Workdept \) is also null for the same tuple (only employee No. 20 in this example).

Note that the difference between \( Q_1 \) and \( Q_2 \) naturally lies on the predicate to be evaluated, but also on the number of tuple variables required. The predicate of \( Q_1 \) is evaluated on pairs of tuples, while \( Q_2 \) considers tuples individually.

Example 3. The following query \( Q_3 \) restricts the scope of \( Q_1 \), leading to the notion of conditional functional dependencies [5]. For example, we consider only employees with a level of qualification above 16.

\[ Q_3: \text{FINDRULES OVER Empno, Lastname, Workdept, Job, Sex, Bonus} \]

\[ \text{SCOPE t1, t2 (SELECT * FROM Emp WHERE Educlevel > 16)} \]

\[ \text{CONDITION ON } \$A \text{ IS t1.$A = t2.$A} \]

Interestingly, \( Sex \rightarrow Bonus \) holds with this restriction, meaning that above a certain level of qualification (16), the gender determines the bonus.

Example 4. Query \( Q_4 \) is an approximation of \( Q_1 \) for numeric values similar to Metric Functional Dependencies [11], where strict equality is discarded to take into account variations under 10%. For instance, salaries 41250 and 38250

Figure 1: Web interface for RQL
are considered close (7.5\% difference), but not salaries 41250 and 36170 (13.1\% difference).

\(Q'_1\): FINDRULES
\[
\text{OVER Educlevel, Sal, Bonus, Comm\\nSCOPE t1, t2 Emp\\nCONDITION ON $A$ IS\\n2*ABS(t1.$A$-t2.$A$)/(t1.$A$+t2.$A$)<0.1}
\]

In that case, \(\text{Sal} \rightarrow \text{Comm}\) holds, meaning that employees earning similar salaries receive similar commissions.

We have shown so far query examples related to implications (in FCA) and functional dependencies (in DB). Nevertheless, RQL is not restricted to these types of queries at all and can express many more rules.

Example 5. Assume we are interested in a kind of sequential dependencies \([9]\), i.e. dependencies showing similar behavior of attribute values. \(Q_3\) discovers numerical attributes that vary together (i.e., \(X \rightarrow Y\) means that if \(X\) increases then \(Y\) also increases).

\(Q_3\): FINDRULES
\[
\text{OVER Educlevel, Sal, Bonus, Comm\\nSCOPE t1, t2 Emp\\nCONDITION ON $A$ IS\\n$\text{Sal} \rightarrow \text{Comm}$ and $\text{Comm} \rightarrow \text{Sal}$ hold in Emp, which means that a higher salary is equivalent to a higher commission.}
\]

Example 6. Continuing the previous example, assume now the analyst wants to focus on male employees (see also Figure 1).

\(Q'_3\): FINDRULES
\[
\text{OVER Educlevel, Sal, Bonus, Comm\\nSCOPE t1, t2 Emp\\nWHERE Sex='M'\\nCONDITION ON $A$ IS $t1.$A > t2.$A}
\]

In that case, \(\text{Educlevel} \rightarrow \text{Bonus}\) also holds, which means that male employees with higher education levels receive higher bonuses.

Example 7. Instead of narrowing the scope of a query, user-defined conditions can bind different tuple variables together with a custom relationship specified in the WHERE clause of the RQL query. For example, \(Q_4\) finds disparities between managers and managees, i.e. rules on attributes for which managers have values greater than or equal to their managees.

\(Q_4\): FINDRULES
\[
\text{OVER Educlevel, Sal, Bonus, Comm\\nSCOPE t1, t2 Emp\\nWHERE t1.Empno = t2.Mgrno\\nCONDITION ON $A$ IS $t1.$A >= t2.$A}
\]

In this example, \(\emptyset \rightarrow \text{Bonus}\) holds in Emp, which means that managers always earn a bonus greater than or equal to their managees'.

3. FEEDBACK THROUGH COUNTEREXAMPLES

Given a RQL query, the data analyst may also interact with the system to know whether or not a given rule holds. She can provide a rule to the system and two cases arise: either the rule holds and the analyst is notified that the rule is indeed valid; or the rule does not hold which means that at least one counterexample exists and one of them is provided by the system. This notion of counterexample is well known for functional dependencies, and provides very good feedback to the data analyst with her own data. We strongly believe that counterexamples are a great tool to help the analyst understand why a particular rule does not hold, and refine if necessary her analysis in an iterative process.

To illustrate counterexamples, suppose that a data analyst wants to explore her hypothesis that higher salaries and higher education levels yield higher bonuses \((\text{Salary, Educlevel} \rightarrow \text{Bonus})\), using \(Q_3\). Figure 4 gives an overview of what RQL provides as a counterexample for this rule, that is, two tuples among which one (employee No. 10) has a higher Salary and Educlevel than the other (employee No. 50), but not a higher Bonus.

With this counterexample as a starting point, and especially the SQL query generated to extract it from the database, the data analyst can quickly switch to SQL to get an idea of why this rule is not verified. For instance, employees that are either female or have a finance job are easily pointed out as having a higher salary and education level, but lower bonuses than others. The data analyst can then refine her RQL query, for example by narrowing the scope of the data, such as in \(Q'_5\) where a higher Educlevel by itself implies a higher Bonus.

The number of tuples required to provide a counterexample depends on the number of tuple variables. \(Q_3\) or FDs need at least two tuples. But with \(Q_2\), one tuple (for instance employee No. 30) is enough to prove that the rule \(\text{Workdept} \rightarrow \text{Mgno}\) does not hold.

4. IMPLEMENTATION AND APPLICATION
The RQL web application has been implemented in Java with the Play Framework [15]. External tools have been used for the most expensive part of the rule generation process, i.e. the enumeration of minimal transversal of hypergraphs [12]. The chosen DBMS is Oracle 11g Release 2.

Importantly, the web interface provides the SQL code generated by the system as often as possible, for instance to identify the counterexample, so that the analyst can issue her own query to identify more counterexamples.

RQL can be interacted with in two modes: (i) a Sample DB is provided with selected examples to offer a quick way into RQL (ii) a Sandbox allows users to upload and query their own data (currently limited to 3 tables and 200 kB). RQL has been used by 120 undergraduate students to play with functional dependencies and other constraints in a database course at INSA Lyon. Not surprisingly, the notion of counterexamples has been widely used by students. RQL has been appreciated for its ability to bridge the gap between the SQL language and functional dependencies in database design by providing a unified interface for both SQL and RQL queries.

Previous works [6] have highlighted the efficiency of RQL as a two step process, even on large databases.

### 5. CONCLUSION

RQL is introduced as a web interface to discover rule patterns over relational databases. RQL subsumes SQL statements by providing the opportunity to specify and get results as a set of rules or some counterexamples. The rule mining problem is seen as a query processing problem, for which we have proposed a query rewriting technique allowing the delegation of as much processing as possible to the underlying DBMS engine [6]. RQL allows SQL developers to extract precise information without any specific knowledge in data mining.

### 6. REFERENCES


---

**Rule verification:**
The rule 
\[
\text{Sal Educlevel bonus}
\]
is ... 
ELSE
ELSE
0
0
0. 
AND
AND
rownum rownum <= 1

---

**Counter-example:**

EMPNO in data mining.
eextract precise information without any specific knowledge underlying DBMS engine [6]. RQL allows SQL developers to mining problem is seen as a query processing problem, for results as a set of rules or some counterexamples. The rule statements by providing the opportunity to specify and get results as a set of rules or some counterexamples. The rule mining problem is seen as a query processing problem, for which we have proposed a query rewriting technique allowing the delegation of as much processing as possible to the underlying DBMS engine [6]. RQL allows SQL developers to extract precise information without any specific knowledge in data mining.

---

**Generated query:**

```sql
SELECT * FROM Emp t1, t2, t3 WHERE t1.Sal + t2.Sal + t3.Sal >= t1.Educlevel + t2.Bonus
AND CASE WHEN (t1.Sal + t2.Sal) > t3.Bonus THEN (t3.Educlevel - 1) ELSE 0 END = 0
```

---

**Figure 4: Counterexample with RQL**