Querying Key-Value Stores Under Simple Semantic Constraints: Rewriting and Parallelization

Olivier Rodriguez, Corentin Colomier, Cecilie Rivière, Reza Akbarinia, Federico Ulliana

To cite this version:

Querying Key-Value Stores Under Simple Semantic Constraints: Rewriting and Parallelization

Olivier Rodriguez
Université de Montpellier
olivier.rodriguez@etu.umontpellier.fr

Corentin Colomier
Université de Montpellier
corentin.colomier@etu.umontpellier.fr

Cecile Riviere
Université de Montpellier
cecilie.riviere@etu.umontpellier.fr

Reza Akbarinia
INRIA Sophia Antipolis
reza.akbarinia@inria.fr

Federico Ulliana
Université de Montpellier
ulliana@lirmm.fr

ABSTRACT

We demonstrate a system for accessing Key-Value stores in the presence of semantic constraints, as considered in the setting of ontology-mediated query answering. The constraints we study are expressed in a native rule language for JSON records and their purpose is to establish a high level view over a collection of legacy Key-Value stores. To ensure correct and complete data access, constraints are taken into account via query rewriting techniques tailored for MongoDB queries. The considered setting also enable the deployment of parallelization techniques for optimizing query rewriting thereby accelerating the whole data access task.

During the demonstration, attendees will be able first to query JSON datasets enriched with semantic constraints. Then, it will be possible to make a detailed analysis of query rewriting algorithm and compare the parallel version with the baseline centralized one.

1 INTRODUCTION

Ontology-mediated query answering (OMQA) is currently at the center of many investigations. This paradigm for data access is designed to query incomplete databases by taking into account semantic constraints which are usually provided by means of an ontology. OMQA has many applications in data integration, especially with Semantic Web technologies [11]. Indeed, constraints can be used to establish a unified view of a collection of heterogenous datasets thereby allowing the user to formulate high-level queries employing a richer vocabulary than that of the single sources.

The paradigm so far has been widely studied for relational and RDF databases. The ontological knowledge is often expressed using description logics [9, 3] which also underpin the OWL 2 QL and EL profiles, or suitable fragments of existential rules (also called Datalog²) akin to Tuple Generating Dependencies (TGDs) [4, 8, 5]. Yet, much less attention has been devoted to the case where the data to be accessed are JSON records sitting in a NOSQL system such as Key-Value stores like MongoDB [2] or CouchDB [1].

To illustrate the OMQA approach to Key-Value stores [6, 7], consider the JSON record $r$ in Figure 1 which describes a computer science university department with two of its professors, and the MongoDB queries $Q_1$ and $Q_2$. The query $Q_1$ selects all department records where there exists a professor with a contact, while the query $Q_2$ selects all computer science departments having a director.

It can be easily seen that these two queries have empty answer when evaluated on $r$. The reason is that they employ a vocabulary which is disjoint from that of $r$. Indeed, $Q_1$ asks for the key contact while $Q_2$ asks for the key director that are both not employed in $r$.

Here is where semantic constraints can come into play. Indeed, although the key contact is not present in the source $r$, this can be seen as a high-level key generalizing phone and mail. This is captured by rules $σ_1$ and $σ_2$. Hence, by taking into account the semantic constraints while accessing data, $r$ would satisfy the query $Q_1$. Similarly, $σ_3$ says that whenever a professor is present, then a director exists. Again, if this constraint is considered, then $r$ would also satisfy $Q_2$. This shows how semantic constraints allow the user to express a high-level query on a collection of data sources and also reason with incomplete information at the record level. Of course, the structure of a record can be enriched by more expressive constraints [6, 10], but in this work we focus on simple constraints built on pairs of keys, as those yet illustrated.

Demo contribution

This demonstration showcases i) a system for accessing Key-Value stores in the presence of semantic constraints based on query rewriting as well as ii) a parallelization technique for optimizing data access. Query rewriting is an algorithmic approach to account for semantics whose goal is to incorporate the constraints into the query. This process yields a set (or union) of rewritings whose answers set over the input database is exactly the same as the initial query under constraints. This virtual approach to query answering has many advantages. First, it is well suited for accessing legacy databases even with read-only access rights. Second, it is resilient to updates, being the rewritings of a query independent from the data sources.

\[
(r) \quad \{ \text{dept} : \{ \text{name} : "CS", \text{prof} : [ ] \} \}
\]

\[
(Q_1) \quad \text{find}(\text{dept} : \{ \text{prof} : [ \text{contact} : [\text{exists} : \text{true}] ] \})
\]

\[
(Q_2) \quad \text{find}(\text{dept} : [ ] \{ \text{contact} : [\text{exists} : \text{true}] ] \})
\]

\[
(σ_1) \quad \text{phone} \rightarrow \text{contact} \quad (σ_2) \quad \text{mail} \rightarrow \text{contact} \quad (σ_3) \quad \text{prof} \rightarrow \exists \text{director}
\]

Figure 1: Record example
The alternative approach to query rewriting is materialization. Intuitively speaking, for the constraints we consider of the form \( k \rightarrow k' \), materialization means to create fresh copies of all values coupled with a key \( k \) and then associating these copies with a new key \( k' \). The tree-shape of JSON records may however hinder compact (graph) representations of data thereby making materialization suboptimal. This is exacerbated by the fact that data is voluminous.

In contrast, queries are usually small. From an OMQA perspective, it is thus interesting to explore query rewriting approaches that can take into account semantic constraints while accessing data without modifying the data sources.

In spite of the many advantages of virtual approaches we already mentioned, query rewriting can still suffer from combinatorial explosions, which often happen in practice. To mitigate this problem, our system features a parallel query rewriting method that can be used to distribute the computing of rewritings over several nodes as well as serve as basis for distributed query evaluation.

Our system currently enables access to MongoDB stores. However, the extension to other Key-Value stores systems is conceptually straightforward. The source of the code used for this demonstration is available at github.com/zuri66/keyval-qrewriting.

## 2 PRELIMINARIES

In this section, we present the preliminaries about the data, queries, rules and query rewriting.

### 2.1 Data

A JSON record, or key-value record, is a finite set of key-value pairs of the form \( r = \{(k_1, e_1), \ldots, (k_n, e_n)\} \). A value is recursively defined as (i) a constant or a null value (ii) a sequence \( e = [e_1, \ldots, e_n] \) where each \( e_i \) is a value and \( n \geq 1 \), or (iii) a record \( r \), where each \( k_i \in \text{Keys} \) and each \( e_i \) is a value, with \( n \geq 1 \) and \( k_i \neq k_j \) whenever \( i \neq j \). A KV-store \( I \) is a set of records.

We see a record \( r \) as a rooted labeled tree, denoted by \( \text{tree}(r) \), in which edges are labeled by keys, leaves are labeled by constant or null values, and all other internal nodes are unlabeled. Of course, a key-value pair \( (k, [e_1, \ldots, e_n]) \) involving a sequence is represented by several edges labeled by \( k \) leading to the nodes that represent the elements of \( e \). Figure 2 provides a tree representation of the record given in the previous example.

\footnote{We do not represent the ordering on the elements of a sequence, since the considered queries will not exploit this order. Moreover, a sequence nested in a sequence is seen as a constant as core queries will not go through them either.}

### 2.2 Queries

We focus on the tree pattern queries that are supported by MongoDB [2]. These are of the form \( \text{find}(\phi) \) where \( \phi \) is an expression generated by the following grammar.

\[
\begin{align*}
\phi & := \{ \text{key : cond} \} \mid \text{conj} \\
\text{cond} & := \phi \mid \text{terminal} \mid \{ \$\text{elemMatch} : [\text{conj}] \} \\
\text{conj} & := \phi \mid \phi \& \phi \mid \phi \mid \phi \\
\text{terminal} & := a \mid a \\
\end{align*}
\]

It is easy to see that the set of queries expressible by this grammar corresponds to a form of tree pattern queries without joins.\(^2\) The grammar should be self explicative, we just comment on the main constructs. The \( \$\text{elemMatch} \) clause is used to specify a conjunction of conditions that the subrecord has to match.\(^3\) The condition \( \exists \text{exists : true} \) is used to verify the existence of any value associated with a key, while \( a \) to specify the constant that has to be matched like the string “CS” for the query \( Q_2 \) in Figure 1. Recall that a KV-store query is always evaluated starting from the root of the record. As for records, we see a query \( \text{find}(\phi) \), as a tree. This is denoted by \( \text{tree}(\phi) \). Figure 3 provides a tree representation of the queries given in the previous example.

Following the MongoDB semantics, a \( \text{find} \) query evaluated on a record \( r \) yields an answer which is either the record \( r \) itself or the emptyset. Accordingly, we say that a record \( r \) answers (or satisfies) a query \( \text{find}(\phi) \) if there exists a (tree-) homomorphism \( h : \text{tree}(\phi) \rightarrow \text{tree}(r) \) mapping the root of \( \text{tree}(\phi) \) to the root of \( \text{tree}(r) \). With a little abuse of notation, we denote this by \( r \models \phi \).\(^4\) For a KV store \( I \), we write \( I \models \phi \) whenever \( r \models \phi \) for some \( r \in I \).

### 2.3 Rules

The constraints we are interested in are key inclusions, as introduced in [10, 6]. These are expressed as rules \( \sigma \) of the form

\[
\begin{align*}
k & \rightarrow k' \quad (\forall\text{-rule}) \\
& \rightarrow \exists k' \quad (\exists\text{-rule})
\end{align*}
\]

and allow to define hierarchies of keys (\( \forall \)-rules) as well as the existence of mandatory keys (\( \exists \)-rule). The semantics of rules is now described. For the sake of formalization, we lift the setting of trees to that of multi-trees. That is, we consider trees where there can be more than one edge between a pair of nodes. To illustrate, \( r = \{(a : (c : \text{"alice"}), b : (c : \text{"alice"})\} \) can be seen either as the tree \( \{(n_0, n_1, a), (n_0, n'_1, b), (n_1, n_2, c), (n'_1, n'_2, c)\} \) where the root is

\footnote{In passing, note also that \( \phi \) is a JSON record.} \footnote{As a subtlety, MongoDB matches only records that belong to a sequence.} \footnote{This stands for \( \text{tree}(r) \equiv \text{tree}(\phi) \).}
We now present an algorithmic procedure for query rewriting. For all $\phi$ the general case \cite{6}, this procedure always terminates, that is, the rewriting process simply replaces the key pair $u$ with $v$ if there exists an edge $(u, v, v')$ in the tree. We denote this by $t \models r \models \phi$. For a set of constraints $\Sigma$ we have that $t \models r \models \phi$ if and only if $t \models r \models \phi$ for all $\phi \in \Sigma$. Then, we say that $t$ is a model of $r \models \phi$, for the set of rewritings in $r \models \phi$. We denote this by $t \models r \models \phi$. For the sake of efficiency, the rewriting load should be well balanced among the computing units. Formally, given a query $\phi$, a set of rules $\Sigma$, and $n$ computing units, our goal is to find a balanced partitioning of the set of rewritings $\{ \Phi_1, \ldots, \Phi_n \}$ making each node $u$ generating a set $\Phi_j$ such that $|\Phi_j| \approx \frac{|\text{Rew}(\phi, \Sigma)|}{n}$. After determining the partitions, each unit starts doing the local rewritings for its partition, and the global result is the union of the local results.

Encoding the space of rewritings. Our solution for load balanced rewriting proceeds in two steps. First, we define a encoding of the space of rewritings suitable for this task. Second, we partition the space over the computing units.

To encode $\text{Rew}(\phi, \Sigma)$ we build a decision tree $\mathcal{G}(\phi, \Sigma)$ such that: i) any path from the root to a leaf of the tree represents a (unique) rewriting of $\phi$ with $\Sigma$; and ii) all rewritings in $\text{Rew}(\phi, \Sigma)$ have a corresponding path in $\mathcal{T}(\phi, \Sigma)$. A balanced partitioning of the rewritings can be obtained by equally partitioning these “rewriting paths” over the $n$ units. We outline the main technical ideas of our approach then generalize it to the whole framework.

Dealing with single rule applications. Consider the query $\phi = \{ \text{faculty} : \{ \text{contact} : \{ \text{sexists : true} \} \} \}$ and the rules $\sigma_1 : \text{phone} \rightarrow \text{contact}$ and $\sigma_2 : \text{prof} \rightarrow \text{faculty}$. In this case, $\text{Rew}(\phi, \{ \sigma_1, \sigma_2 \}) = \{ \text{faculty.contact, faculty.prof, prof.contact, prof.phone} \}$. This set can be represented with a binary tree whose depth $\ell$ depends on the number of rewriteable edges of the query (in this case 2) and where an edge from level $i$ to level $i+1$ has a binary label corresponding to the fact that the edge has been rewritten (1) or not (0) by the respective rule. The absolute path of each leaf of the tree corresponds to a rewriting. With this approach we have at most $2^{|\phi|}$ binary codes, one for each rewriting. Note that $0^\ell$ denotes the original query.

3 PARALLELIZATION

We now present a parallel method for computing $\text{Rew}(\phi, \Sigma)$. The idea is to exploit a set of independent computing units $u_1, \ldots, u_n$ being each a local thread or a machine of a computing cluster. Each unit $u_i$ will compute a (sub)set of the rewritings of the input query $\Phi_i \subseteq \text{Rew}(\phi, \Sigma)$. It is important that every unit computes different rewritings, that is for all $u_i \neq u_j$ we have that $\Phi_i \cap \Phi_j = \emptyset$.

For the sake of efficiency, the rewriting load should be well balanced among the computing units. Formally, given a query $\phi$, a set of rules $\Sigma$, and $n$ computing units, our goal is to find a balanced partitioning of the set of rewritings $\{ \Phi_1, \ldots, \Phi_n \}$ making each node $u_i$ generating a set $\Phi_i$ such that $|\Phi_i| \approx \frac{|\text{Rew}(\phi, \Sigma)|}{n}$. After determining the partitions, each unit starts doing the local rewritings for its partition, and the global result is the union of the local results.

Encoding the space of rewritings. Our solution for load balanced rewriting proceeds in two steps. First, we define a encoding of the space of rewritings suitable for this task. Second, we partition the space over the computing units.

To encode $\text{Rew}(\phi, \Sigma)$ we build a decision tree $\mathcal{T}(\phi, \Sigma)$ such that: i) any path from the root to a leaf of the tree represents a (unique) rewriting of $\phi$ with $\Sigma$; and ii) all rewritings in $\text{Rew}(\phi, \Sigma)$ have a corresponding path in $\mathcal{T}(\phi, \Sigma)$. A balanced partitioning of the rewritings can be obtained by equally partitioning these “rewriting paths” over the $n$ units. We outline the main technical ideas of our approach then generalize it to the whole framework.

Dealing with single rule applications. Consider the query $\phi = \{ \text{faculty} : \{ \text{contact} : \{ \text{sexists : true} \} \} \}$ and the rules $\sigma_1 : \text{phone} \rightarrow \text{contact}$ and $\sigma_2 : \text{prof} \rightarrow \text{faculty}$. In this case, $\text{Rew}(\phi, \{ \sigma_1, \sigma_2 \}) = \{ \text{faculty.contact, faculty.prof, prof.contact, prof.phone} \}$. This set can be represented with a binary tree whose depth $\ell$ depends on the number of rewriteable edges of the query (in this case 2) and where an edge from level $i$ to level $i+1$ has a binary label corresponding to the fact that the edge has been rewritten (1) or not (0) by the respective rule. The absolute path of each leaf of the tree corresponds to a rewriting. With this approach we have at most $2^{|\phi|}$ binary codes, one for each rewriting. Note that $0^\ell$ denotes the original query.
the sake of efficiency, it is important to have a bounded communication between \( u \) and each other unit.

An important property of our encoding approach is that the obtained codes can be put in bijection with the integers \([0, \ldots, k-1]\). Therefore, the unit \( u \) can communicate to each unit \( u_i \) in a succinct way the interval of rewritings that it has to generate. In fact, \( u \) sends to \( u_i \) only two values \((N, N + \lambda)\) where \( N \) is a positive integer and \( \lambda \approx k/n \). For instance, if \( n = 1 \) the interval \([0,3]\) representing the set of (binary) codes \([00,01,10,11]\) will be sent to \( u_i \). This ensures a \( O(1) \) bound for communicating the interval to a single computing unit.

**Dealing with multiple rule applications.** In the general case, the rule application is more complex than what we saw in the previous example. Since not only two rules \( \sigma_1 \) and \( \sigma_2 \) can rewrite the same edge of the query, but also the application of \( \sigma_1 \) can enable that of \( \sigma_2 \). For these cases, the approach yet described must be properly extended. To do so, we look at transitive dependencies between the keys in \( \Sigma \).

We say that \( k' \) is \((3)\)-derivable from \( k \), denoted by \( k \leq_3 k' \), if there exists a sequence of rules \( \sigma_1, \ldots, \sigma_n \) of the form \( \sigma_i = k_i \rightarrow \{3\}k_i' \) such that \( k_i' = k_{i+1} \) for all \( 1 \leq i < n \), with \( k_1 = k \) and \( k_n = k' \). When all rules in the sequence are \((V)\)-rules we say that \( k' \) is \((V)\)-derivable from \( k \), and we denote it by \( k \leq V k' \). It follows that \( k \leq V k' \) implies \( k \leq_3 k' \), but the converse does not hold. It is important to distinguish the two relations because, as we mentioned, \((3)\)-rules must be applied only on the existential leaves of the query. Finally, note that we can compute the relations \( \leq_V \) and \( \leq_3 \) in a preprocessing step with a single traversal of \( \Sigma \).

Let us denote by the integers \([1, \ldots, \rho]\) the edges of the query \( \phi \) that can be rewritten by a rule of \( \Sigma \). Then, each edge \( j \) labelled with \( k' \), is rewritten with a key \( k \) such that i) \( k \leq_3 k' \) if the edge is incident to an existential leaf and ii) \( k \leq V k' \) otherwise. Importantly, we denote by \( b_j \) the number of ways that the edge \( j \) of the query can be rewritten. For instance, from the example of Figure 1 we have phone \( \leq_V \) contact and mail \( \leq_Y \) contact and thus for an edge \( j \) of the query labelled with contact we have \( b_j = 2 \).

The set \( \text{Rew}(\phi, \Sigma) \) can therefore be represented by an unranked decision tree where i) each level \( j \) corresponds to the rewritings of the edge \( j \) of the query, and ii) the degree of all of nodes at level \( j \) is exactly \( (b_j + 1) \), that is, there exists an outgoing edge from the node for each possible rewriting of the edge (plus one denoting that the edge is not rewritten). As before, the leaves of the decision tree are labelled with their absolute path \( c_1 c_2 \cdots c_{\rho} \) which describe a unique rewriting of \( \phi \). Note that \( c_j \leq (b_j + 1) \) for all \( 1 \leq j \leq \rho \).

To preserve the bounded-size communication when moving from binary to unranked trees, we need to put again the set of \( k \) codes we obtained in bijection with the integers \([0, \ldots, k-1]\), in order to easily compute the partitioning of the rewritings over the \( n \) computing units. To do so, we see a code as a number in a multiple base \((b_p, \ldots, b_1)\) where as before each \( b_j \) denotes the number of elements of the base (corresponding to the number of possible rewritings of the edge \( j \) in the query). A code \((c_1 \ldots c_{\rho})\) in the base \((b_p, \ldots, b_1)\) is equivalent to the integer \( p \) where

\[
p = c_p \ast B_1 + c_{p-1} \ast B_2 + \ldots + c_1 \ast B_p
\]

where \( B_1 = 1 \) and for all \( i \geq 2 \) it holds that \( B_i = b_{i-1} \ast B_{i-1} \). In the above example, the codes \([00, 01, 02, 10, 11, 12]\) will correspond to the interval \([0, 5]\). Indeed, we consider each code in the base\(^6\) \((b_2, b_1) = (2, 3)\) we have \( B_1 = 1 \) (by definition) and \( B_2 = b_1 \ast B_1 = 3 \). For example, the code \((c_1 c_2) = (12)\) stands for \( c_2 \ast B_1 + c_1 \ast B_2 = 5 \). At this point, by knowing the base \((b_n, \ldots, b_1)\) it is possible to determine the number of rewritings \( k = b_p \ast B_1 + \cdots + b_1 \ast B_p \), and then assign to each of the \( n \) units an interval of size \( k/n \).

### 4 SYSTEM ARCHITECTURE

In this section, we describe the global functioning of our system, whose architecture is depicted in Figure 4.

Our solution can be implemented in any shared nothing parallel framework. In our demo, we use different cores of a machine and parallelize our solution by executing threads in the cores. However, the threads can also be executed in the nodes of a distributed cluster, if such a cluster is available.

The first component of our system is a **Parser** of the queries and the rules. After that, the **Code Generator** is in charge of computing the encoding of the rewritings. The Code Generator associates to every edge of the query, which can be rewritten, all of the possible keys and their associated codes for the rewriting. This module is also in charge of properly taking into account the \((3)\)-rule constraints which, as said before, must apply only to existential leaves. The set of codings is then partitioned into intervals. Each

---

\(^6\)we have two possibilities for \( c_j \) (i.e., 0 and 1) and tree for \( c_j \) (i.e., 0, 1, 2)
task (in a unit) takes its interval, and generates the corresponding rewriting results. When all parallel rewriting tasks are completed, the results are combined in order to obtain the final rewriting set. Finally, the Data Access module is in charge of evaluating the whole query set on the available MongoDB instances.

5 DEMO OVERVIEW
The goal of our demonstration is i) to showcase the benefits of accessing data under semantic constraints as well as ii) to show the efficiency of the parallelization for the query rewriting system.

The demonstration scenario is constituted of two JSON datasets. The first, University, has been manually built and concerns university employees, faculty members, and students. The second, XMark [12], is a standard benchmark for semi-structured data that we used to generate JSON records. Both datasets are associated with queries of different complexities ranging from path queries to tree-pattern queries. Similarly semantic constraints adapted to the structure of the datasets have been devised.

The demonstration will start by showing how thanks to semantic constraints we can bring novel and pertinent answer to users queries as well as formulate high-level queries over heterogenous records. Then, it will be possible to analyze the functioning of the query rewriting module of the system, by exploring the rewriting set generated for a given input query and constraint set. Finally, it will be shown how, for cases where the number of rewriting is huge, the rewriting tasks can be parallelized across multiple threads so as to optimise the overall data access.

The demonstration will take place on a graphical interface allowing the attendees to choose predefined queries and rules, but also to write their own queries and constraints. Similarly, the number of rewriting threads can be configured so as to compare the effectiveness of the parallelized version of the algorithm with the baseline centralized one.

REFERENCES