Efficient OLAP Operations For RDF Analytics
Elham Akbari-Azirani, François Goasdoué, Ioana Manolescu, Alexandra Roatis

To cite this version:

HAL Id: hal-01187448
https://hal.inria.fr/hal-01187448
Submitted on 28 Aug 2015

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.
Efficient OLAP Operations For RDF Analytics

Elham Akbari Azirani#1, François Goasdoué∗#2, Ioana Manolescu #∗#3, Alexandra Roatiş #∗#4

# INRIA, France  * U. Rennes 1, France  * U. Paris-Sud, France
1elham@lri.fr  2fg@irisa.fr  3ioana.manolescu@inria.fr  4roatis@lri.fr

ABSTRACT

RDF is the leading data model for the Semantic Web, and dedicated query languages such as SPARQL 1.1, featuring in particular aggregation, allow extracting information from RDF graphs. A framework for analytical processing of RDF data was introduced in [1], where analytical schemas and analytical queries (cubes) are fully re-designed for heterogeneous, semantic-rich RDF graphs. In this novel analytical setting, we consider the following optimization problem: how to reuse the materialized result of a given RDF analytical query (cube) in order to compute the answer to another cube. We provide view-based rewriting algorithms for these cube transformations, and demonstrate experimentally their practical interest.

1. MOTIVATION AND OUTLINE

Graph-structured, semantics-rich, heterogeneous RDF data needs dedicated warehousing tools [2]. In [1], we have introduced a novel framework for RDF data analytics. At its core lies the concept of analytical schema (AnS, in short), which reflects a view of the data under analysis. Based on an analytical schema, analytical queries (AnQ, in short) can be posed over the data; they are the counterpart of the “cube” queries in the relational data warehouse scenario, but specific to our RDF context. Just like in the traditional relational data warehouse (DW) setting, new algorithms are needed in our context, where, un-

2. BACKGROUND: RDF ANALYTICS

RDF analytical schemas and queries. RDF analytical schemas can be seen as lenses through which data is analyzed. An AnS is a labeled directed graph, whose nodes are analysis classes and whose edges are analysis properties, deemed interesting by the data for a specific analysis task. The instance of an AnS is built from the base data; it is an RDF graph itself, heterogeneous and semantic-rich, restructured for the needs of the analysis.

Figure 1 shows a sample AnS for analyzing bloggers and blog posts. An AnS node is defined by an unary query, which, evaluated over an RDF graph, returns a set of URIs. For instance, the node Blogger is defined by a query which (in this example) returns the URIs user1, user2 and user3. The interpretation is that the AnS defines the analysis class Blogger, whose instances are these three users. An AnS edge is defined by a binary query, returning pairs of URIs from the base data. The interpretation is that for each (a, o) URI pair returned by the query defining the analysis property p, the triple a p o holds, i.e., o is a value of the property p of a. Crucial for the ability of AnSs to support analysis of heterogeneous RDF graphs is the fact that AnS nodes and edges are defined by completely independent queries. Thus, for instance, a user may be part of the Blogger instance whether or not the RDF graph comprises value(s) for the analysis properties identifiedBy, livesIn etc. of that user. Further, just like in a regular RDF graph, each blogger may have multiple values for a given analysis property. For instance, user1 is identified both as William and as Bill.

We consider the conjunctive subset of SPARQL consisting of basic graph pattern (BGP) queries, denoted $q(\bar{x}) : t_1, \ldots, t_n$, where \{$t_1, \ldots, t_n$\} are triple patterns. Unless we explicitly specify that a query has bag semantics, the default semantics we consider is that of set.

The head of $q$, denoted head($q$) is $q(\bar{x})$, while the body $t_1, \ldots, t_n$ is denoted body($q$). We use letters in italics (possibly with subscripts) to denote variables. A rooted BGP query is a query $q$ where each variable is reachable through triples from a distinguished variable, denoted root. For instance, the following query is a rooted BGP, whose root is $t_1$: 

![Figure 1: Sample Analytical Schema (AnS).](image-url)
\[ q(x_1, x_2, x_3) := x_1 \mathit{acquaintedWith} x_2, \]
\[ x_1 \mathit{identifiedBy} y_1, \]
\[ x_1 \mathit{wrotePost} y_2, y_2 \mathit{postedOn} x_3 \]

The query’s graph representation below shows that every node is reachable from the root \( x_1 \).

\[ \begin{align*}
\mathit{acquaintedWith} & \quad \rightarrow \quad x_2 \\
\mathit{identifiedBy} & \quad \rightarrow \quad y_1 \\
\mathit{wrotePost} & \quad \rightarrow \quad y_2 \
\end{align*} \]

An analytical query consists of two BGP queries homomorphic to the \( \mathit{Ans} \) and rooted in the same \( \mathit{Ans} \) node, and an aggregation function. The first query, called a classifier \( \mathcal{C} \), specifies the facts and the aggregation dimensions, while the second query, called the measure, returns the values that will be aggregated for each fact. The measure query has bag semantics. Example 1 presents an \( \mathit{AnQ} \) over the \( \mathit{Ans} \) defined in Figure 1.

**Example 1. (Analytical Query)** The query below asks for the number of sites where each blogger posts, classified by the blogger’s age and city:

\[ Q := \mathcal{C}(x, d_{\mathit{age}}, d_{\mathit{city}}), m(x, v_{\mathit{site}}), \mathit{count} \]

where the classifier and measure queries are defined by:

\[ \mathcal{C}(x, d_{\mathit{age}}, d_{\mathit{city}}) := x \mathit{rdf:type Blogger}, \]
\[ x \mathit{hasAge} d_{\mathit{age}}, x \mathit{livesIn} d_{\mathit{city}} \]
\[ m(x, v_{\mathit{site}}) := x \mathit{rdf:type Blogger}, \]
\[ x \mathit{wrotePost} p, p \mathit{postedOn} v_{\mathit{site}}. \]

The semantics of an analytical query is:

**Definition 1. (Answer Set of an Analytical Query)** Let \( \mathcal{I} \) be the instance of an analytical schema with respect to some RDF graph. Let \( Q := \mathcal{C}(x, d_1, \ldots, d_n), m(x, v), \oplus \) be an analytical query against \( \mathcal{I} \).

The answer set of \( Q \) against \( \mathcal{I} \), denoted \( \mathit{ans}(Q, \mathcal{I}) \), is:

\[ \mathit{ans}(Q, \mathcal{I}) = \{ (d_1', \ldots, d_n', \oplus(q'(\mathcal{I}))) | \{ x_1', \ldots, x_k' \} \subseteq \mathcal{C}(\mathcal{I}) \text{ and } q'(\mathcal{I}) \text{ is the bag of all values } v_{j}' \text{ such that } \{ x_1', \ldots, x_k', v_j' \} \in m(\mathcal{I}) \} \]

where \( q'(\mathcal{I}) \) is a bag containing all measure values \( v \) corresponding to \( x \), and the operator \( \oplus \) aggregates all members of this bag. If \( q'(\mathcal{I}) \) is empty, the aggregated measure is undefined, and \( x \) does not contribute to the cube. In the following, for conciseness, we use \( \mathit{ans}(Q, \mathcal{I}) \) to denote \( \mathit{ans}(Q, \mathcal{I}) \), where \( \mathcal{I} \) is considered the working instance of the analytical schema.

In other words, the \( \mathit{AnQ} \) returns each tuple of dimension values from the answer of the classifier query, together with the aggregated result of the measure query for those dimension values. The answer set of an \( \mathit{AnQ} \) can thus be represented as a cube of \( n \) dimensions, holding in each cube cell the corresponding aggregate measure.

The counterpart of a fact, in this framework, is any value to which the first variable in the classifier, \( x \) above, is bound, and that has a non-empty answer for the measure query. In RDF, a resource may have zero, one or several values for a given property. Accordingly, in our framework, a fact may have multiple values for each measure; in particular, some of these values may be identical, yet they should not be collapsed into one. For instance, if a product is rated by 5 users, one of which rates it 3 while the four others rate it 4, the number of times each value was recorded is important. This is why we assign bag semantics to \( q'(\mathcal{I}) \). In all other contexts, we consider BGP with set semantics; this holds in particular for any classifier query \( c(x, d_1, \ldots, d_n) \).

**Example 2. (Analytical Query Answer)** Consider the \( \mathit{AnQ} \) in Example 1 over the \( \mathit{Ans} \) in Figure 1. Suppose the classifier query’s answer set is:

\[ \{ \{ \mathit{user}_1, 28, \mathit{Madrid} \}, \{ \mathit{user}_2, 35, \mathit{NY} \}, \{ \mathit{user}_3, 35, \mathit{NY} \} \} \]

the measure query is evaluated for each of the three facts, leading to the intermediary results:

\[ x^2\]
\[ q'(\mathcal{I}) \]
\[ \langle \{s_1\}, \{s_2\} \rangle \]
\[ \langle \{s_3\} \rangle \]

where \( \{ \} \) denotes the bag constructor. Aggregating the sites among the classification dimensions leads to the \( \mathit{Ans} \) answer:

\[ \{ \{ 28, \mathit{Madrid}, 3 \}, \{ 35, \mathit{NY}, 2 \} \} \]

**OLAP for RDF.** On-Line Analytical Processing (OLAP) technology enhances the abilities of data warehouses (so far, mostly relational) to answer multi-dimensional analytical queries. In a relational setting, the so-called “OLAP operations” allow computing a cube (the answer to an analytical query) out of another previously materialized cube.

In our data warehouse framework specifically designed for graph-structured, heterogeneous RDF data, a cube corresponds to an \( \mathit{AnQ} \); for instance, the query in Example 1 models a bi-dimensional cube on the warehouse related to our sample \( \mathit{Ans} \) in Figure 1. Thus, we model traditional OLAP operations on cubes as \( \mathit{AnQ} \) rewritings, or more specifically, rewritings of extended \( \mathit{AnQs} \) which we introduce below.

**Definition 2. (Extended \( \mathit{AnQ} \))** Let \( S \) be an \( \mathit{Ans} \), and \( d_1, \ldots, d_n \) be a set of dimensions, each ranging over a non-empty finite set \( V_i \). Let \( \Sigma \) be a total function over \( \{ d_1, \ldots, d_n \} \) associating to each \( d_i \), either \( V_i \) or a non-empty subset of \( V_i \). An extended analytical query \( Q \) is defined by a triple:

\[ Q := c_\Sigma(x, d_1, \ldots, d_n), m(x, v), \oplus \]

where \( c \) is a classifier and \( m \) a measure query over \( S \), \( \oplus \) is an aggregation operator, and moreover:

\[ c_\Sigma(x, d_1, \ldots, d_n) = \bigcup \{ x \mid x \in \Sigma(d_i) \} \]

In the above, the extended classifier \( c_\Sigma(x, d_1, \ldots, d_n) \) is the set of all possible classifiers obtained by replacing each dimension \( d_i \) with a value from \( \Sigma(d_i) \). We introduce \( \Sigma \) to constrain some classifier dimensions, i.e., to restrict the classifier result. The semantics of an extended analytical query is derived from the semantics of a standard \( \mathit{AnQ} \) (Definition 1) by replacing the tuples from \( c(\mathcal{I}) \) with tuples from \( c_\Sigma(\mathcal{I}) \). Thus, an extended analytical query can be seen as a union of a set of standard \( \mathit{AnQs} \), one for each combination of values in \( \Sigma(d_1), \ldots, \Sigma(d_n) \). Conversely, an analytical query corresponds to an extended analytical query where \( \Sigma \) only contains pairs of the form \( (d_i, V_i) \).

We define the following RDF OLAP operations:
A SLICE operation binds an aggregation dimension to a single value. Given an extended query \( Q = (c_{\Sigma} (x, d_1, \ldots, d_n), m(x, v)), \) a SLICE operation over a dimension \( d_i \) with value \( v_i \) returns the extended query \( Q_{\text{SLICE}} = (c_{\Sigma} (x, d_1, \ldots, d_{i-1}, d_{i+1}, \ldots, d_n), m(x, v_i)) \), where \( \Sigma' = (\Sigma \setminus \{d_i\}) \cup \{(d_i, v_i)\} \).

Similarly, a DICE operation constrains several aggregation dimensions to values from specific sets. A DICE on \( Q \) over dimensions \( \{d_{i_1}, \ldots, d_{i_k}\} \) and corresponding sets of values \( \{S_{i_1}, \ldots, S_{i_k}\} \), returns the query \( Q_{\text{DICE}} = (c_{\Sigma} (x, d_{i_1}, \ldots, d_{i_k}), m(x, v), \oplus) \), where \( \Sigma' = (\Sigma \setminus \{d_{i_1}, \ldots, d_{i_k}\}) \cup \{(d_{i_1}, S_{i_1}), \ldots, (d_{i_k}, S_{i_k})\} \).

A DRILL-OUT operation on \( Q \) over dimensions \( \{d_{i_1}, \ldots, d_{i_k}\} \) corresponds to removing these dimensions from the classifier. It leads to a new query \( Q_{\text{DRILL-OUT}} \), having the classifier \( c_{\Sigma'} (x, d_{j_1}, \ldots, d_{j_{n-k}}) \), where \( d_{j_1}, \ldots, d_{j_{n-k}} \in \{d_{i_1}, \ldots, d_{i_k}\} \setminus \{d_{i_1}, \ldots, d_{i_k}\} \) and \( \Sigma' = (\Sigma \setminus \{d_{i_1}, \ldots, d_{i_k}\}) \).

Finally, a DRILL-IN operation on \( Q \) over dimensions \( \{d_{n+1}, \ldots, d_{n+k}\} \) which all appear in the classifier’s body and have value sets \( \{V_{n+1}, \ldots, V_{n+k}\} \) corresponds to adding these dimensions to the head of the classifier. It produces a new query \( Q_{\text{DRILL-IN}} \) having the classifier \( c_{\Sigma'} (x, d_{n+1}, \ldots, d_{n+k}) \), where the dimensions \( d_{n+1}, \ldots, d_{n+k} \notin \{d_{i_1}, \ldots, d_{i_k}\} \), \( \Sigma' = (\Sigma \setminus \{d_{i_1}, \ldots, d_{i_k}\}) \).

These operations are illustrated in the following example.

**EXAMPLE 3. (OLAP OPERATIONS)** Let \( Q \) be the extended query corresponding to the query-cube defined in Example 7 that is: \( Q = (c(x, d_{age}, d_{city}), m(x, v_{site}), \text{count}) \), \( \Sigma = \{(d_{age}, V_{age}), (d_{city}, V_{city})\} \) (the classifier and measure are as in Example 7).

A SLICE operation on the \( d_{age} \) dimension with value \( 35 \) replaces the extended classifier of \( Q \) with \( c_{\Sigma'} (x, d_{age}, d_{city}) = \{c(x, 35, d_{city})\} \) where \( \Sigma' = \Sigma \setminus \{(d_{age}, V_{age})\} \cup \{(d_{age}, \{35\})\} \).

A DICE operation on both \( d_{age} \) and \( d_{city} \) dimensions with values \( \{28\} \) for \( d_{age} \) and \{Madrid, Kyoto\} for \( d_{city} \) replaces the extended classifier of \( Q \) with \( c_{\Sigma'} (x, d_{age}, d_{city}) = \{c(x, 28, \text{Madrid}), c(x, 28, \text{Kyoto})\} \) where \( \Sigma' = \{(d_{age}, \{28\}), (d_{city}, \{\text{Madrid, Kyoto}\})\} \).

A DRILL-OUT on the \( d_{age} \) dimension produces \( Q_{\text{DRILL-OUT}} = (c_{\Sigma'} (x, d_{age}), m(x, v_{site}), \text{count}) \) with \( \Sigma' = \{(d_{age}, V_{age}), (d_{city}, V_{city})\} \) and \( \text{body}(c') \equiv \text{body}(c) \).

Finally, a DRILL-IN on the \( d_{age} \) dimension applied to the query \( Q_{\text{DRILL-OUT}} \) above produces \( Q \), the query of Example 7.

**3. OPTIMIZED OLAP OPERATIONS**

The above OLAP operations lead to new queries, whose answers can be computed based on the \( AnS \) instance. The focus of the present work is on answering such queries by using the materialized results of the initial \( AnQ \), and (only when that input is insufficient) more data, such as intermediary results generated while computing \( AnQ \) results, or (a small part of) the \( AnS \) instance. These results are often significantly smaller than the full instance, hence obtaining the answer to the new query based on them is likely faster than computing it from the instance. Figure 3 provides a sketch of the problem.

In the following, all relational algebra operators are assumed to have bag semantics.

Given an analytical query \( Q \) whose measure query (with bag semantics) is \( m \), we denote by \( \hat{m} \) the set-semantics query whose body is the same as the one of \( m \) and whose head comprises all the variables of \( m \)’s body. Obviously, there is a bijection between the bag result of \( m \) and the set result of \( \hat{m} \). Using \( \hat{m} \), we define next the intermediary answer of an \( AnQ \).

**DE**

**3. (INTERMEDIARY QUERY OF AN ANQ)** Let \( Q \) :-

![Figure 2: Problem statement.](image-url)

It is easy to see that \( \text{int}(Q) \) holds all the information needed in order to compute both \( c \) and \( m \); it holds more information than the results of \( c \) and \( m \), given that it preserves all the different embeddings of the (bag-semantics) \( m \) query in the data. Clearly, evaluating \( \text{int} \) is at least as expensive as evaluating \( Q \) itself, while \( \text{int} \) is conceptually useful, we do not need to evaluate it or store its results.

Instead, we propose to evaluate (possibly as part of the effort for evaluating \( Q \)) store and reuse a more compact result, defined as follows. For a given query \( Q \) whose measure (with bag semantics) is \( m \), we term extended measure result over an instance \( I \), denoted \( m^\epsilon(I) \), the set defined by:

\[
\{\text{new}(k, t) \mid t \in m(I)\}
\]

where \( \text{new}(k, t) \) is a key-creating function returning a distinct value at each call. A very simple implementation of \( \text{new}(k, t) \), which we will use for illustration, returns successively 1, 2, 3, etc. We assign a key to each tuple in the measure so that multiple identical values of a given measure for a given fact would not be erroneously collapsed into one. For instance, if

\[
m(I) = \{(x_1, m_1), (x_1, m_1), (x_1, m_2), (x_2, m_3)\},
\]

then

\[
m^\epsilon(I) = \{(1, x_1, m_1), (2, x_1, m_1), (3, x_1, m_2), (4, x_2, m_3)\}.
\]

**DEF**

**4. (PARTIAL RESULT OF AN ANQ)** Let \( Q \) :- \((c, m, \oplus)\) be an \( AnQ \). The partial result of \( Q \) on an instance \( I \), denoted \( \text{pres}(Q, I) \) is:

\[
\text{pres}(Q, I) = c(I) \bowtie_{e} m^\epsilon(I)
\]

One can see \( \text{pres}(Q, I) \) as the input to the last aggregation performed in order to answer the \( AnQ \), augmented with a key. In the
following, we use \( \text{pres}(Q) \) to denote \( \text{pres}(Q, I) \) for the working instance of the AnS.

**Problem Statement:** (Answering AnQs Using the Materialized Results of Other AnQs) Let \( Q, Q_T \) be AnQs such that applying the OLAP transformation \( T \) on \( Q \) leads to \( Q_T \).

The problem of answering \( Q_T \) using the materialized result of \( Q \) consists of finding: (i) an equivalent rewriting of \( Q_T \) based on \( \text{pres}(Q) \) or \( \text{ans}(Q) \), if one exists; (ii) an equivalent rewriting of \( Q_T \) based on \( \text{pres}(Q) \) and the AnS instance, otherwise.

Importantly, the following holds:

\[
\pi_{d_1, \ldots, d_n, v}(\text{int}(Q)(I)) = \pi_{d_1, \ldots, d_n, v}(\text{pres}(Q, I)) \quad (1)
\]

\[
\forall_{d_1, \ldots, d_n, v} \pi_{d_1, \ldots, d_n, v}((\text{int}(Q))) \quad (2)
\]

\[
\text{ans}(Q)(I) = \forall_{d_1, \ldots, d_n, v} \pi_{d_1, \ldots, d_n, v}(\text{pres}(Q, I)) \quad (3)
\]

Equation (1) directly follows from the definition of \( \text{pres} \). Equation (2) will be exploited to establish the correctness of some of our techniques. Equation (3) above is the one on which our rewriting-based AnQ answering technique is based.

### 3.1 Slice and Dice

In the case of slice and dice operations, the data cube transformation is made simply by row selection over the materialized final results of an AnQ.

**Example 4.** (DICE) The query \( Q \) asks for the average number of words in blog posts, for each blogger's age and residential city.

\[
Q ::= (c(x, d_{age}, d_{city}), m(x, v_{words}), \text{average})
\]

\[
c(x, d_{age}, d_{city}) ::= \text{x rdf:type blogger,}
\]

\[
m(x, v_{words}) ::= \text{x hasAge d_{age}, x livesIn d_{city},}
\]

\[
p hasWordCount v_{words}
\]

Suppose the answer of \( c \) over \( I \) is

\[
\{(\text{user}, 28, \text{Madrid}), (\text{user}, 35, \text{NY}), (\text{user}, 28, \text{Madrid})\}
\]

and the answer of \( m \) over \( I \) is

\[
\{(\text{user}, 100), (\text{user}, 120), (\text{user}, 35, 570), (\text{user}, 410)\}
\]

Joining the answers of \( c \) and \( m \) in such a query results in:

\[
\{(\text{user}, 28, \text{Madrid}), (\text{user}, 28, \text{Madrid}), (\text{user}, 28, \text{Madrid})\}
\]

The final answer to \( Q \) after aggregation is:

\[
\{(28, \text{Madrid}, 210), (35, \text{NY}, 570)\}
\]

The query \( \text{Q\_dice} \) is the result of a DICE operation on \( Q \), restricting the \( d_{age} \) to values between 20 and 30. \( \text{Q\_dice} \) differs from \( Q \) only by its classifier which can be written as \( c_{28}(x, d_{age}, d_{city}) \) where \( \Sigma' = \Sigma \setminus \{d_{age}, V_{age}\} \cup \{(d_{age}, \{d_{age} \mid 20 \leq d_{age} \leq 30\})\} \).

Applying DICE on the answer to \( Q \) above yields the result:

\[
\{(28, \text{Madrid}, 210)\}
\]

Now, we calculate the answer to \( \text{Q\_dice} \). The result of the classifier query \( c_{\gamma} \), obtained by applying a selection on the \( d_{age} \) dimension is:

\[
\{(\text{user}, 28, \text{Madrid}), (\text{user}, 28, \text{Madrid})\}
\]

Evaluating \( m \) and joining its result with the above set yields:

\[
\{(\text{user}, 28, \text{Madrid}, 100), (\text{user}, 28, \text{Madrid}, 120), (\text{user}, 28, \text{Madrid}, 410)\}
\]

The final answer to \( \text{Q\_dice} \) after aggregation is:

\[
\{(28, \text{Madrid}, 210)\}
\]

**DICE applied over the answer of \( Q \) yields the answer of \( \text{Q\_dice} \)**

**Definition 5.** (Selection) Let \( \text{dice} \) be a dice operation on analytical queries. Let \( \Sigma' \) be the function introduced in Definition 2. We define a selection \( \text{slice}(\text{dice}) \) as a function on the space of analytical query answers \( \text{ans}(Q) \) where:

\[
\text{slice}(\text{ans}(Q)) = \{(d_1, \ldots, d_n, v)|\langle d_1, \ldots, d_n, v \rangle \in \text{ans}(Q) \quad \forall i \in \{1, \ldots, n\} \quad \exists d_i \in \Sigma'(d_i)\}
\]

**Proposition 1.** Let \( Q ::= (c_{\Sigma}(x, d_1, \ldots, d_n), m(x, v, \oplus)) \) and \( Q_{\_\text{dice}} ::= (c_{\Sigma'}(x, d_1, \ldots, d_n), m(x, v, \oplus)) \) be two analytical queries such that query \( Q_{\_\text{dice}} = \text{dice}(Q) \). Then \( \text{slice}(\text{ans}(Q)) = \text{ans}(Q_{\_\text{dice}}) \).

The proofs for all our results can be found in 3.2 Drill-Out

### 3.2 Drill-Out

Unlike the relational DW setting, in our RDF warehousing framework the result of a drill-out operation (that is, the answer to \( \text{Q\_drill-out} \)) cannot be computed directly from the answer to the original query \( Q \), and here is why. Each tuple in \( \text{ans}(Q) \) binds a set of dimension values to an aggregated measure. In fact, each such tuple represents a set of facts having the same dimension values. Projecting a dimension out will make some of these sets merge into one another, requiring a new aggregation of the measure values. Computing this new aggregated measure from the ones in \( \text{ans}(Q) \) will require considering whether the aggregation function has the distributive property, i.e., whether \( \oplus(a, \oplus(b, c)) = \oplus(a, \oplus(b, c)) \).

1. **Distributive aggregation function**, e.g. \( \text{sum} \). In this case, the new aggregated measure value could be computed from \( \text{ans}(Q) \) if the sets of facts aggregated in each tuple of \( \text{ans}(Q) \) were mutually exclusive. This is not the case in our setting where each fact can have several values along the same dimension. Thus, aggregating the already aggregated measure values will lead to erroneously consider some facts more than once; avoiding this requires being able to trace the measure results back to the facts they correspond to.

2. **Non-distributive aggregation function**, e.g. \( \text{avg} \). For such functions, the new aggregated measure must be computed from scratch.

Based on the above discussion, we propose Algorithm 1 to compute the answer to \( \text{Q\_drill-out} \), using the partial result of \( Q \), denoted \( \text{pres}(Q) \) above, which we assume has been materialized and stored as part of the evaluation of the original query \( Q \). This deduplication (\( D \)) step is needed, since some facts may have been repeated in \( T \) for being multivalued along \( d_j \). The aggregation function \( \oplus \) is applied to the measure column of the resulting relation \( T \), using \( \gamma \), grouping the tuples along the dimensions \( d_1, \ldots, d_{\gamma - 1}, d_{\gamma + 1}, \ldots, d_n \).
Algorithm 1 DRILL-OUT cube transformation

1: Input: \( \text{pres}(Q), d_i \)
2: \( T \leftarrow \Pi_{\text{root},d_i,d_{i+1},...,d_n} (\text{pres}(Q)) \)
3: \( T \leftarrow \delta(T) \)
4: \( T \leftarrow \gamma_{d_1,...,d_{i-1},d_{i+1},...,d_n} (\text{pres}(Q)) \)
5: return \( T \)

Example 5. (DRILL-OUT) Consider an analytical query \( Q \) such that its classifier \( C_1 \), measure \( m \) and intermediary answer \( \text{ans}(Q) \) have the results shown below:

\[
\begin{array}{c|cc|c|c}
\text{root} & d_1 & \ldots & d_{n-1} & d_n & m \\
\hline
x & a_1 & \ldots & a_{n-1} & a_n & m_1 \\
y & a_1 & \ldots & a_{n-1} & b_n & m_2 \\
\end{array}
\]

Let \( \text{pres}(Q) \) be the result of a DRILL-OUT operation on \( Q \) eliminating dimension \( d_n \). The measure of \( \text{pres}(Q) \) is still \( m \), while its classifier \( C_2 \) has the answer shown next:

\[
\begin{array}{c|c|c}
\text{root} & d_1 & \ldots & d_{n-1} \\
\hline
x & a_1 & \ldots & a_{n-1} & 1 & m_1 \\
y & a_1 & \ldots & a_{n-1} & 2 & m_2 \\
\end{array}
\]

Note that \( C_2 \) returns only one row for \( x \), because it has one value for the dimension vector \( \langle d_1, \ldots, d_{n-1} \rangle \): \( \text{pres}(Q)_{\text{DRILL-OUT}} \) yields:

\[
\begin{array}{c|cc|c|c}
\text{root} & d_1 & \ldots & d_{n-1} & k & v \\
\hline
x & a_1 & \ldots & a_{n-1} & 1 & m_1 \\
y & a_1 & \ldots & a_{n-1} & 2 & m_2 \\
\end{array}
\]

Applying aggregation over the above table leads to:

\[
\begin{array}{c|cc}
d_1 & \ldots & d_{n-1} & v \\
\hline
a_1 & \ldots & a_{n-1} & \oplus \{m_1, m_2\} \\
\end{array}
\]

Algorithm 2 uses the partial result of \( \text{Q} \), denoted \( \text{pres}(Q) \), and consults the materialized \( \text{Ans} \) instance to obtain the missing information necessary to answer \( \text{Q}_{\text{DRILL-OUT}} \). We retrieve this information through an auxiliary query defined as follows.

**Definition 6. (Auxiliary DRILL-IN Query)** Let \( Q : \langle \text{c}(x, d_1, \ldots, d_n, m(x, v), \oplus) \rangle \) be an \( \text{Ans} \) and \( \text{d}_{n+1} \) a non-distinguished variable in \( c \). The auxiliary DRILL-IN query of \( Q \) over \( \text{d}_{n+1} \) is a conjunctive query \( q_{\text{aux}}(\text{dvars}, \text{d}_{n+1}) : \text{body}_{\text{aux}} \) where:

- each triple \( t \in \text{body}_c \) containing the variable \( \text{d}_{n+1} \) is also in \( \text{body}_{\text{aux}} \).
Based on $I$, the answer to $Q_{\text{drill-in}}^Q$ is:

<table>
<thead>
<tr>
<th>x</th>
<th>d₂</th>
<th>d₃</th>
<th>v</th>
</tr>
</thead>
<tbody>
<tr>
<td>video1</td>
<td>URL1</td>
<td>firefox</td>
<td>n</td>
</tr>
<tr>
<td>video1</td>
<td>URL2</td>
<td>chrome</td>
<td>n</td>
</tr>
</tbody>
</table>

Joining the above with $\text{pres}(Q)$ yields the last table in Figure 3 which after aggregation yields the result of $Q_{\text{drill-in}}$.

4. RELATED WORK

Previous RDF data management research focused on efficient stores, query processing, view selection etc. BGP query answering techniques have been studied intensively, e.g., [3][5][6], and some are deployed in commercial systems such as Oracle 11g’s “Semantic Graph” extension. Our optimizations can be deployed on top of any RDF data management platform, to extend it with optimized analytic capabilities.

The techniques we presented can be seen as a particular case of view-based rewriting [7], where partial AnQ results are used as a materialized view. Novel algorithms were required due to the novel AnQ language we introduced in [1].

SPARQL 1.1 [8] features SQL-style grouping and aggregation, less expressive than our AnQs, as our measure queries allow more flexibility than SPARQL. Thus, the OLAP operation optimizations we presented can also apply to the more restricted SPARQL analytical context.

OLAP has been thoroughly studied in a relational setting, where it is at the basis of a successful industry; in particular, OLAP operation evaluation by reusing previous cube results is well-known. The heterogeneity of RDF, which in turn justified our novel RDF analytics framework [1], leads to the need for the novel algorithms we described here, which are specific to this setting.

5. CONCLUSION

Our work focused on optimizing the OLAP transformations in the RDF data warehousing framework we introduced in [1], by using view-based rewriting techniques. To this end, for each OLAP operation, we introduced an algorithm that answers a transformed query based on the final or on an intermediary result of the original analytical query. We formally prove the correctness of our techniques, and describe experiments performed with our algorithms, in our technical report [3].

6. REFERENCES


