A Framework for Efficient Representative Summarization of RDF Graphs
Šejla Čebirić, François Goasdoué, Ioana Manolescu

To cite this version:

HAL Id: hal-01577778
https://inria.hal.science/hal-01577778
Submitted on 28 Aug 2017

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.
A Framework for Efficient Representative Summarization of RDF Graphs

Šejla Ćebirić\textsuperscript{1} François Goasdoué\textsuperscript{2,1}
Ioana Manolescu\textsuperscript{1}
\textsuperscript{1}INRIA and Ecole Polytechnique, France \textsuperscript{2}Univ. Rennes 1, France
August 28, 2017

Abstract

RDF is the data model of choice for Semantic Web applications. RDF graphs are often large and have heterogeneous, complex structure. Graph summaries are compact structures computed from the input graph; they are typically used to simplify users’ experience and to speed up graph processing.

We introduce a formal RDF summarization framework, based on graph quotients and RDF node equivalence; our framework can be instantiated with many such equivalence relations. We show that our summaries represent the structure and semantics of the input graph, and establish a sufficient condition on the RDF equivalence relation which ensures that a graph can be summarized more efficiently, without materializing its implicit triples.

1 Introduction

To facilitate working with very large, complex-structure, heterogeneous graphs, many graph summaries have been proposed, including some specifically tailored for Resource Description Framework (RDF) graphs \cite{1,5,6}. A summary of an RDF graph $G$ is a smaller graph (typically also RDF), based on which questions about $G$ may be answered more efficiently than by using $G$ directly.

In this work, we define a formal generic summarization framework for RDF graphs, based on the classical notion of graph quotients, and on our notion of RDF node equivalence. While quotient-style summaries have been studied in the past \cite{1,3}, our first contribution is a formal framework for summarizing RDF graphs including possible RDF Schema constraints, which leads us to study the interplay between summarization and saturation with such constraints. Specifically, our second contribution is a sufficient condition on the RDF node equivalence relation, which guarantees that the summary of the saturation of $G$ can be built through a shortcut procedure, without actually saturating $G$; this can significantly speed up the summary construction.

Our summaries, representative of the complete (saturated) graphs but often much smaller, can be used in GUIs to help users explore and query RDF graphs, or to
An RDF graph is a set of triples \((s, p, o)\) where \(s\) is termed the subject, \(p\) the property, and \(o\) the object; such a triple states that \(s\) has the property \(p\) that has value \(o\). Well-formed triples, as per the RDF specification, belong to \((U \cup B) \times U \times (U \cup L \cup B)\), where \(U\) is a set of Universal Resource Identifiers (URIs in short), \(L\) is a set of literals (constants), and \(B\) is a set of blank nodes, representing unknown URI or literal values. A triple \((s, p, o)\) states that its subject \(s\) has the property \(p\) whose value is the object \(o\). RDF allows making class assertions, if \(p\) is the special built-in RDF property rdf:type (\(τ\) in short), and property assertions otherwise (Table 1).

### RDF Schema statements (at the bottom left of Table 1) together with the shorthand notations of their properties allow specifying ontological constraints relating classes and/or properties. The semantics of an RDF graph \(G\) is its saturation (or closure) \(G^∞\), defined as the \(G\) triples together with all the implicit triples that can be derived from them and the entailment rules from the RDF standard. Table 1 (right) shows rules that use RDFS constraints to derive implicit facts or implicit constraints. Figure 1 depicts a sample publications graph, where \(Pub\) stands for publication (\(CPub\) in conferences and \(JPub\) in journals), \( JPub\) in journals, and \( JPub\) in books. Figure 1 shows the entailment rule for the property assertion \((rdfs:prefix,τ,o)\) and the shorthand notations for the entailment rules from the RDF standard. Table 1 shows rules that use RDFS constraints to derive implicit facts or implicit constraints.
in journals), a for author, t for title and y for year; class nodes and RDFS triples appear in blue, for instance, the domain of t (title) is Pub. Solid arrows correspond to explicit G triples, and dotted arrows to implicit triples; all together, they depict $G^\infty$.

3 Summarization framework

We start by recalling the classical notion of graph quotient. Let $G = (V, E)$ be a labeled directed graph whose edges $E$ have labels from a set $A$. Let $\sim$ be an equivalence relation over the graph node set $V$. The quotient of $G$ through $\sim$, denoted $G/\sim$, is a labeled directed graph having (i) a node $n_S$ for each set $S$ of equivalent $V$ nodes, and (ii) an edge $n_{S_1} \xrightarrow{a} n_{S_2}$ for some label $a \in A$ iff there exist two $V$ nodes $n_1 \in S_1$ and $n_2 \in S_2$ such that the edge $n_1 \xrightarrow{a} n_2 \in E$.

When summarizing an RDF graph, class and schema information (e.g., the blue part of Figure 1) should be preserved, as they encode its semantics. Thus, we define:

**Definition** Let $\equiv$ be a binary relation between the nodes of an RDF graph. We say $\equiv$ is an RDF equivalence relation iff (i) $\equiv$ is reflexive, symmetric and transitive, (ii) any class node is $\equiv$ only to itself, and (iii) any property node is $\equiv$ only to itself.

We define an RDF summary as a graph quotient w.r.t. a given RDF node equivalence:

**Definition** Given an RDF graph $G$ and an RDF node equivalence relation $\equiv$, the summary of $G$ by $\equiv$, which is an RDF graph denoted $G/\equiv$, is the quotient of $G$ by $\equiv$.

Different RDF equivalence relations lead to different summaries. Figures 2 illustrates two of them on the saturated $G^\infty$ from Figure 1. Circles denote new-URI summary nodes, each of which represents a set of $G$ nodes. For instance, at left, a single node represents $r_1, r_2, r_3$; at right, they are separated by their sets of types. Below, we do not discuss any particular summary further; instead, we focus on our summarization framework, and its interplay with saturation.

For a summary to reflect (represent) a graph $G$, queries having answers on $G$ should also have answers on the summary. Given an RDF query language $Q$, we define:

**Definition** Let $G$ be any RDF graph. $G/\equiv$ is $Q$-representative of $G$ if and only if for any query $q \in Q$ such that $q(G) \neq \emptyset$, we have $q((G/\equiv)^\infty) \neq \emptyset$.

We target summaries representative of any query over the graph structure of $G$, including imprecise queries using variables in some property positions. Thus, we instantiate $Q$ into Extended Relational Basic Graph Pattern Queries (RBGP*, in short), a core fragment of SPARQL, defined as follows. A query triple pattern belongs to $V \times (U \cup V) \times V$ or $V \times \{\tau\} \times U$, where $V$ is a set of variables. An RBGP* query $q$ is of the
form: \( q(\bar{x}) \leftarrow t_1, \ldots, t_n \) where each \( t_i \) is a query triple pattern, \( \{t_1, \ldots, t_n\} \) is noted \textit{body}(q), and \( \bar{x} \), called the \textit{answer variables}, is a subset of the variables in \textit{body}(q). A sample RBGP* query is: \( q^*(x_1, x_3) \leftarrow (x_1, \tau, \text{Book}), (x_1, \text{author}, x_2), x_2 \ y \ x_3 \).

We show (the proof is in [2]) that for any RDF equivalence relation \( \equiv \):

\textbf{Proposition 3.1} An RDF summary \( G_{/\equiv} \) is RBGP*-representative.

RBGP* representativeness ensures that any query specifying a certain graph pattern in \( G \) and/or querying the structure itself (by means of variables in property positions, such as \( y \) in the sample query above) which has answers on \( G \), also has answers on \( G_{/\equiv} \).

In the presence of an RDF Schema, the semantics of \( G \) is its saturation \( G^\infty \). Thus, a representative summary must reflect both the explicit and the implicit triples of \( G \). For instance, the summaries in Figure 2 show that some \( G^\infty \) resources (e.g., \( r_1, r_2, r_3 \)) are of type \( \text{Pub} \), but the same summaries computed from \( G \) alone do not, as the corresponding \( \tau \) triples are implicit in \( G \). A simple way to obtain \( (G^\infty)_{/\equiv} \) is to compute \( G^\infty \) and then summarize it. We define a novel alternative \textit{shortcut} method, which avoids saturating \( G \), yet it constructs an RDF graph \textit{strongly isomorphic} to \( (G^\infty)_{/\equiv} \), as follows:

\textbf{Definition} A \textit{strong isomorphism} between two RDF graphs \( G_1, G_2 \), noted \( G_1 \equiv G_2 \), is an isomorphism which is the identity for the class and property nodes.

\textbf{Definition} Summarization through the RDF node equivalence relation \( \equiv \) \textit{admits a shortcut} iff for any RDF graph \( G \), \( (G^\infty)_{/\equiv} \equiv ((G_{/\equiv})^\infty)_{/\equiv} \) holds.

The shortcut summarizes \( G \), saturates the result, then summarize it again (the three green edges in Figure 3). Its result is \textit{essentially} \( (G^\infty)_{/\equiv} \), as the two have identical graph structures (guaranteed by the strong isomorphism), on which RBGP* representativeness is defined. They only differ in the new URIs of their nodes (circles in Figure 2).

What is the interest of the shortcut? If \( G_{/\equiv} \) is much smaller than \( G \), it is much faster to saturate \( G_{/\equiv} \) (on the shortcut) than to saturate \( G \); \( (G_{/\equiv})^\infty \) is also likely to be small, thus fast to summarize. Further, summarizing \( G^\infty \) is faster than summarizing \( G \), given that \( G^\infty \) is at least as large as \( G \). Summing up these inequalities, the \textit{time spent on the shortcut may be (much) shorter than the time spent to build} \( (G^\infty)_{/\equiv} \) \textit{directly}.

By the summary definition, to every node in \( G \) corresponds exactly one node in the summary \( G_{/\equiv} \). We call \textit{representation function} and denote \( f_{/\equiv} \) the function associating a summary node to each \( G \) node; we say \( f_{/\equiv}(n) \) \textit{represents} \( n \) in the summary. An important structural property relates \( G, G^\infty \) and the function \( f_{/\equiv} \) (see Figure 3):

\textbf{Lemma 3.2 (Summarization Homomorphism)} Let \( G \) be an RDF graph, \( G_{/\equiv} \) its summary and \( f_{/\equiv} \) the corresponding representation function from \( G \) nodes to \( G_{/\equiv} \) nodes. \( f_{/\equiv} \) defines a homomorphism from \( G^\infty \) to \( (G_{/\equiv})^\infty \).

Based on the Lemma, we establish the sufficient condition [2] (see Figure 3):

\textbf{Theorem 3.3 (Sufficient condition for shortcuts)} Given an RDF node equivalence relation \( \equiv \), and an RDF graph \( G \), let \( G_{/\equiv} \) be its summary and \( f_{/\equiv} \) the corresponding representation function from \( G \) nodes to \( G_{/\equiv} \) nodes.

If \( \equiv \) satisfies: for any RDF graph \( G \) and any pair \( (n_1, n_2) \) of \( G \) nodes, \( n_1 \equiv n_2 \) in \( G^\infty \) iff \( f_{/\equiv}(n_1) \equiv f_{/\equiv}(n_2) \) in \( (G_{/\equiv})^\infty \), then \( (G^\infty)_{/\equiv} \equiv ((G_{/\equiv})^\infty)_{/\equiv} \) holds.
The summary illustrated at left Figure 3 turns out to admit the shortcut; in our experiments, the shortcut was up to 20x faster than saturating $G$ and then summarizing $G^\infty$. The summary illustrated at right in Figure 3 does not admit the shortcut.

**Conclusion and perspectives** Finding a necessary (and sufficient) condition for the shortcut is currently open. We have instantiated our framework into many summaries, and implemented a summarization tool available online (together with many sample summaries) at [https://team.inria.fr/cedar/projects/rdfsummary](https://team.inria.fr/cedar/projects/rdfsummary). We are currently working on summary-based query pruning, where we decide based on $(G^\infty)/\equiv$ whether a query may have answers on $G^\infty$ or not.

**References**


