A Framework for Efficient Representative Summarization of RDF Graphs

Šejla Čebirić[1], François Goasdoué[2,1], and Ioana Manolescu[1]

[1] INRIA, France
[2] Univ. Rennes 1, France

Abstract. RDF is the data model of choice for Semantic Web applications, and it is often used to model large and heterogeneous dataset. We consider the problem of determining as quickly as possible if a query q lacks answers on a given RDF graph $\mathtt{G}$: from a user perspective, this allows disregarding graphs uninteresting for them; from a query processing perspective, this saves system resources as a query which is known to lack answers does not need to be evaluated on $\mathtt{G}$.
To address this, we introduce a generic *RDF summarization framework* whereas from an RDF $\mathtt{G}$, we build a summary $\mathtt{G}_{/\equiv}$, also an RDF graph but often many orders of magnitude smaller, and such that *if q lacks answers on $\mathtt{G}$, it is guaranteed to lack answers on $\mathtt{G}_{/\equiv}$*. Further, in the presence of RDF Schema, the interesting question is whether q has answers on the *saturation* of $\mathtt{G}$, denoted $\mathtt{G}^{\infty}$. One could address that by saturating $\mathtt{G}$ (if not already done), then computing $(\mathtt{G}^{\infty})_{/\equiv}$ and evaluating q on it. For more efficiency, we introduce a *shortcut procedure* by which we compute the summary of $(\mathtt{G}^{\infty})$ *directly from* $\mathtt{G}$ without saturating it, and provide a sufficient condition for an RDF summary defined according to our framework to admit this shortcut.

1 Summarization framework

This section recalls the RDF summarization framework defined in [**?**,**?**] on which we build the present paper.

Definition 1. *RDF node equivalence Let $\equiv$ be a binary relation between the nodes of an RDF graph. We say $\equiv$ is an* RDF node equivalence relation *(or* RDF equivalence*, in short) iff (i) $\equiv$ is an equivalence relation in the classical sense (it is reflexive, symmetric and transitive), (ii) any class node is $\equiv$ only to itself, and (iii) any property node is $\equiv$ only to itself.*

An RDF summary is defined as a graph quotient with respect to a given RDF node equivalence:

Definition 2. *RDF summary Given an RDF graph $\mathtt{G}$ and an RDF node equivalence relation $\equiv$, the* summary of $\mathtt{G}$ by $\equiv$*, which is an RDF graph denoted $\mathtt{G}_{/\equiv}$, is the quotient of $\mathtt{G}$ by $\equiv$. $\mathtt{G}_{/\equiv}$ data nodes use fresh URIs to identify the sets of equivalent $\mathtt{G}$ data nodes.*

Importantly, any RDF summary enjoys the following property:

Proposition 1. *Schema preservation An RDF graph* $\mathtt{G}$ *and an RDF summary* $\mathtt{G}_{/\equiv}$ *of it have the same schema triples, i.e.,* $\mathtt{S}_{\mathtt{G}} = \mathtt{S}_{\mathtt{G}_{/\equiv}}$ *holds.*

For a summary to reflect (*represent*) the input graph, queries having answers on $\mathtt{G}$ should also have answers on the summary. Given an *RDF query language (dialect)* $\mathcal{Q}$, we define:

Definition 3. *Query-based representativeness Let* $\mathtt{G}$ *be any RDF graph.* $\mathtt{G}_{/\equiv}$ *is* $\mathcal{Q}$-representative of $\mathtt{G}$ *if and only if for any query* $q \in \mathcal{Q}$ *such that* $q(\mathtt{G}^{\infty}) \neq \emptyset$*, we have* $q((\mathtt{G}_{/\equiv})^{\infty}) \neq \emptyset$*.*

Representativeness is desirable as the summary can be used to help users formulate queries: therefore, it is important to reflect all graph patterns that may occur in the data.

We focus on the following query language:

Definition 4. *RBGP* queries An* extended relational *(RBGP*, in short) query is a BGP query whose body has* (i) *URIs or variables in all the property positions,* (ii) *a URI in the object position of every* τ *triple, and* (iii) *variables in any other positions.*

⟹ Both languages [SC: "Both" refers to RBGP and RBGP* but RBGP definition is missing here] forbid URIs or literals in subject and object positions, and require that if type triples are specified in the query, the type is known. They differ in that RBGPs require URIs in the property positions, whereas RBGP* also allow variables there. Clearly, RBGPs are a restriction of RBGP*. A sample RBGP* query is:

$$q^*(x_1, x_3) \text{ :- } x_1\ \tau\ \mathtt{Book},\ x_1\ \mathtt{author}\ x_2,\ x_2\ y\ x_3$$

We define *RBGP* representativeness* by instantiating $\mathcal{Q}$ in Definition 3 to RBGP* queries (Definition 4).

Based on the above definitions, we established [**?**,**?**]:

Proposition 2. *Summary representativeness An RDF summary* $\mathtt{G}_{/\equiv}$ *is RBGP*-representative.*

Given that the semantics of $\mathtt{G}$ is $\mathtt{G}^{\infty}$, an RBGP* representative summary must reflect both the explicit and the implicit triples of $\mathtt{G}$. A straightforward way to obtain $(\mathtt{G}^{\infty})_{/\equiv}$ is to compute $\mathtt{G}^{\infty}$ and then summarize it. This is not directly possible when one does not have the right to add triples to $\mathtt{G}$, and when it is possible, it may be time and space-consuming. Further, it has to be maintained when data or schema triples in $\mathtt{G}$ change.

To avoid going through the saturation step, we identify a method called *shortcut* (to be defined shortly), which guarantees that we can build RBGP* representative summaries efficiently. It allows constructing an RDF graph *strongly isomorphic* to $(\mathtt{G}^{\infty})_{/\equiv}$, which turns out to be *also* a summary of $\mathtt{G}^{\infty}$, as strongly isomorphic graphs are identical up to renaming of their *data* node URIs:

Definition 5. *Strong isomorphism* $\hat{=}$ *A* strong isomorphism between two RDF graphs $\mathtt{G}_1, \mathtt{G}_2$*, noted* $\mathtt{G}_1 \hat{=} \mathtt{G}_2$*, is an isomorphism which is the identity for the class and property nodes.*

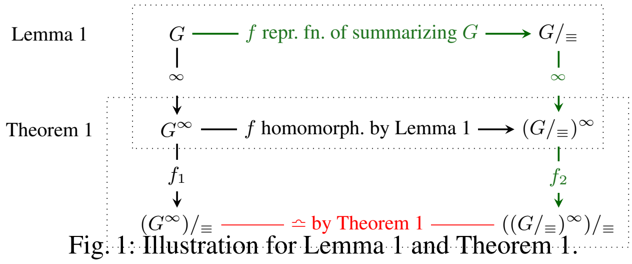

Fig. 1: Illustration for Lemma 1 and Theorem 1.

Definition 6. *Shortcut Summarization through the RDF node equivalence relation* $\equiv$ *admits a* shortcut *iff for any RDF graph* $\mathtt{G}$*,* $(\mathtt{G}^\infty)_{/\equiv} \simeq ((\mathtt{G}_{/\equiv})^\infty)_{/\equiv}$ *holds, where* $\simeq$ *denotes a strong isomorphism (Definition 5).*

The efficient method (or shortcut) to build $(\mathtt{G}^\infty)_{/\equiv}$ introduced above is: summarize $\mathtt{G}$; saturate the result; then summarize it again. The shortcut leads to a graph whose saturation is strongly isomorphic to that of $(\mathtt{G}^\infty)_{/\equiv}$. The two may differ in the exact URIs of the summary *data* nodes (depicted as blank circles in this paper's examples); these are just *representatives* of $\mathtt{G}$ data node groups, and their exact URIs do not matter. In contrast, the summary structure is of crucial importance as representativeness relies on it; this structure is preserved by $\simeq$. Thus, essentially, *a shortcut allows to obtain (an equivalent to) the summary of the saturated* $\mathtt{G}$ *without saturating it*.

The next theorem establishes a sufficient condition on an RDF node equivalence relation for the existence of a shortcut. To be able to state the theorem, we introduce a new concept and a Lemma.

Representation function f By the summary definition, to every node in $\mathtt{G}$ corresponds exactly one node in the summary $\mathtt{G}_\equiv$. We call *representation function* and denote $f_\equiv$ (or simply f, when this does not cause confusion) the function associating a summary node to each $\mathtt{G}$ node; we say $f(n)$ *represents* n in the summary. An important structural property relates $\mathtt{G}$, $\mathtt{G}^\infty$ and the function f:

Lemma 1 (Summarization Homomorphism). *Let* $\mathtt{G}$ *be an RDF graph,* $\mathtt{G}_{/\equiv}$ *its summary and* f *the corresponding representation function from* $\mathtt{G}$ *nodes to* $\mathtt{G}_{/\equiv}$ *nodes.* f *defines a homomorphism from* $\mathtt{G}^\infty$ *to* $(\mathtt{G}_{/\equiv})^\infty$.

Theorem 1 (Existence of shortcuts). *Given an RDF node equivalence relation* $\equiv$*, and an RDF graph* $\mathtt{G}$*, let* $\mathtt{G}_{/\equiv}$ *be its summary and* $f_\equiv$ *the corresponding representation function from* $\mathtt{G}$ *nodes to* $\mathtt{G}_{/\equiv}$ *nodes.*

If $\equiv$ *satisfies: for any RDF graph* $\mathtt{G}$ *and any pair* (n_1, n_2) *of* $\mathtt{G}$ *nodes,* $n_1 \equiv n_2$ *in* $\mathtt{G}^\infty$ *iff* $f(n_1) \equiv f(n_2)$ *in* $(\mathtt{G}_{/\equiv})^\infty$*, then* $(\mathtt{G}^\infty)_{/\equiv} \simeq ((\mathtt{G}_{/\equiv})^\infty)_{/\equiv}$ *holds.*

Importantly, *not all RDF equivalence relations admit a shortcut*, as we will illustrate in Section **??**.

The condition identified in Theorem 1 is sufficient; finding a necessary (and sufficient) condition is currently open.