GDE: General Data Exchange with Schema and Data Level Mappings

Rana Awada and Iluju Kiringa

University of Ottawa, SITE
rawad049@uottawa.ca kiringa@site.uottawa.ca

1 Introduction

Data exchange (DE) [5, 3] and data coordination [1, 2, 6] are two important settings that were introduced previously in the literature to resolve the problem of integrating information that resides in different sources. A DE setting moves data residing in independent applications, which refer to the same object using the same name, and accesses it through a new target schema. However, a data coordination setting allows the access of data residing in independent sources and that possibly belong to different sets of vocabularies, without necessarily exchanging it and while maintaining autonomy.

Although a data coordination setting provides users with an amalgamated view of related information, this solution is not enough for applications that require a view of related information using a unified set of vocabularies for periodic reporting and decision making. We introduce a *general data exchange* (GDE) setting that extends DE settings to allow collaboration at the instance level, using a mapping table M, that specifies for each constant value in the source, the set of related (or corresponding) constant values in the target.[1]

We show in this paper that a GDE setting can be formalized using the knowledge exchange framework introduced in [4]. It allows us to store a target knowledge base (KB) which consists of a subset of the explicit data exchanged that is necessary to infer the full set of exchanged information using a set Σ_t of FO sentences. We identify in our work the class of "best" KBs to materialize and we define the set of certain answers.

2 Preliminaries

A (DE) *setting* [5, 3] is a tuple $\mathfrak{S} = (\mathbf{S}, \mathbf{T}, \Sigma_{st})$, where $\mathbf{S}$ is a source schema, $\mathbf{T}$ is a target schema, $\mathbf{S}$ and $\mathbf{T}$ do not have predicate symbols in common, and Σ_{st} consists of a set of *source-to-target tuple-generating* dependencies (st-tgds) that establish the relationship between source and target schemas. A st-tgd is a FO-sentence of the form: $\forall \bar{x} \forall \bar{y} \, (\phi(\bar{x}, \bar{y}) \rightarrow \exists \bar{z} \, \psi(\bar{x}, \bar{z}))$, where $\phi(\bar{x}, \bar{y})$ and $\psi(\bar{x}, \bar{z})$ are conjunctions of relational atoms over $\mathbf{S}$ and $\mathbf{T}$ respectively. Let Const and Var be infinite and disjoint sets of constants and nulls, respectively. We consider in our

[1] We consider in this work a particular interpretation of related data in a mapping table; that is, a source element is always uniquely identified by at least one target element.

work "complete" source instances I of $\mathbf{S}$, where it holds that $dom(I) \subseteq \mathsf{Const}$ and do not contain missing data in the form of nulls. However, a target instance J of $\mathbf{T}$, is allowed to contain null values, and it holds that $dom(J) \subseteq \mathsf{Const} \cup \mathsf{Var}$;

A knowledge base [4] over a schema $\mathbf{R}$ is a pair (K, Σ), where K is an instance of $\mathbf{R}$ (the explicit data) and Σ is a set of logical sentences over $\mathbf{R}$ (the implicit data). The set of *models* of (K, Σ), denoted by $\mathsf{Mod}(K, \Sigma)$, is defined as the set of instances of $\mathbf{R}$ that contain the explicit data in K and the implicit data in Σ; that is, $\mathsf{Mod}(K, \Sigma)$ corresponds to the set $\{K' \mid K'$ is an instance of $\mathbf{R}$, $K \subseteq K'$ and $K' \models \Sigma\}$. From now on, $K_{\mathbf{R}'}$ denotes the restriction of instance K to a subset $\mathbf{R}'$ of its schema $\mathbf{R}$.

Mapping tables [6] are mechanisms that establish how values from different domains correspond. In its simplest form, given two domains D_1 and D_2, not necessarily disjoint, a mapping table over $(\mathsf{D}_1, \mathsf{D}_2)$ is a subset of $\mathsf{D}_1 \times \mathsf{D}_2$. Let $\mathsf{Const}^{\mathbf{S}}$ and $\mathsf{Const}^{\mathbf{T}}$ be the sets of source and target constants respectively. We consider in our work mapping tables with the following property: for each value $a \in \mathsf{Const}^{\mathbf{S}} \cap dom(M)$, there exists at least a single target value $a' \in \mathsf{Const}^{\mathbf{T}} \cap dom(M)$ such that $M(a, a')$ holds, and there does not exist a source value $b \in \mathsf{Const}^{\mathbf{S}} \cap dom(M)$, where b is different than a and $M(b, a')$ holds. We say a' uniquely identifies a in M. We define C as the set of values in $dom(M) \cap \mathsf{Const}^{\mathbf{T}}$ that uniquely identify source values mapped in M.

3 GDE a Knowledge Exchange System

A GDE setting $\mathfrak{S} = (\mathbf{S}, \mathbf{T}, \mathcal{M}, \Sigma_{st})$ extends a DE setting with (1) a binary relation symbol $\mathcal{M}$ that appears neither in $\mathbf{S}$ nor in $\mathbf{T}$, and that is called a *source-to-target mapping*; and (2) Σ_{st} that consists of a set of *mapping* st-tgds, which are FO sentences of the form: $\forall\bar{x}\forall\bar{y}\forall\bar{z}\,(\phi(\bar{x}, \bar{y}) \wedge \mu(\bar{x}, \bar{z}) \rightarrow \exists\bar{w}\,\psi(\bar{z}, \bar{w}))$, where (a) $\phi(\bar{x}, \bar{y})$ and $\psi(\bar{z}, \bar{w})$ are conjunctions of relation symbols over $\mathbf{S}$ and $\mathbf{T}$ respectively, and (b) $\mu(\bar{x}, \bar{z})$ is a conjunction of relation symbols that only use the st-mapping relation symbol $\mathcal{M}$. We denote st-mapping tables by M.

In a GDE setting, source KBs are of the form $((I \cup \{M\}), \Sigma_s = \emptyset)$, which correspond to data in the source instance I and the st-mapping table M. On the other hand, the target KBs are of the form $((J \cup \{M\}), \Sigma_t)$ where Σ_t is a set of FO sentences, of type $full$ tgds (which are tgds that do not use existential quantication). We formalize the notion of a (universal) GDE KB-solution, extending the notion of knowledge exchange (universal) solution in [4] to allow coordinating the source and target information provided by M, as follows:

1. J is a *GDE KB-solution* for I and M under $\mathfrak{S}$, if for every $K \in \mathsf{Mod}((J \cup \{M\}), \Sigma_t)$ there is $K' \in \mathsf{Mod}((I \cup \{M\}), \Sigma_s = \emptyset))$ such that the following hold: (a) $K'_M \subseteq K_M$, and (b) $((K'_{\mathbf{S}} \cup K'_M), K_{\mathbf{T}}) \models \Sigma_{st}$.
2. Also, J is a *universal GDE KB-solution* (UGDE) for I and M under $\mathfrak{S}$, if J is a GDE KB-solution, and for every $K' \in \mathsf{Mod}((I \cup \{M\}), \Sigma_s = \emptyset)$ there is $K \in \mathsf{Mod}((J \cup \{M\}, \Sigma_t)$ such that (a) $K_M \subseteq K'_M$, and (b) $((K'_{\mathbf{S}} \cup K'_M), K_{\mathbf{T}}) \models \Sigma_{st}$.

Intuitively, in a GDE setting $\mathfrak{S}$, C is the sole set of target values that can capture correctly the set of source values exchanged to a target instance. There-

fore, intuitively a GDE KB-solution J in $\mathfrak{S}$ has a domain $dom(J) \subseteq C \cup \mathsf{Var}$. We define Σ_t as the following set of full tgds over a schema $\mathbf{T} \cup \{\mathcal{M}, \textsc{Related}\}$, where Related is a fresh binary table:

1. For each $T \in \mathbf{T} \cup \{\mathcal{M}\}$ of arity n and $1 \leq i \leq n$:
   $\forall x_1 \cdots \forall x_n (T(x_1, \ldots, x_i, \ldots, x_n) \rightarrow \textsc{Related}(x_i, x_i))$.
2. $\forall x \forall y \forall z (\mathcal{M}(z, x) \wedge \mathcal{M}(z, y) \wedge C(x) \rightarrow \textsc{Related}(x, y))$.
3. For each $T \in \mathbf{T}$ of arity n:
   $\forall x_1, y_1 \cdots \forall x_n, y_n \, (T(x_1, \ldots, x_n) \wedge \bigwedge_{i=1}^{n} \textsc{Related}(x_i, y_i) \rightarrow T(y_1, \ldots, y_n))$.

In a GDE setting, we define "best" solutions formally following [4] as: Let $\mathfrak{S}$ be a GDE setting, I be a source instance, M an st-mapping table, and J a UGDE solution for I and M under $\mathfrak{S}$. Then J is a *minimal* UGDE solution, if (1) there is no proper subset J' of J such that J' is a UGDE solution for I and M under $\mathfrak{S}$, and (2) there is no UGDE solution J' such that $dom(J') \cap \mathsf{Const}^{\mathbf{T}}$ is properly contained in $dom(J) \cap \mathsf{Const}^{T}$. Also, given a fixed GDE setting, generating UGDE solutions and minimal UGDE solutions is in Logspace.

4 Query Answering

We adapt the notion of a certain answer in the usual DE setting to the GDE setting. Formally, let $\mathfrak{S}$ be a GDE setting, I a source instance, M an st-mapping table, and Q a conjunctive query over $\mathbf{T}$. The set of certain answers of Q over I and M and under $\mathfrak{S}$, denoted $\mathsf{certain}_{\mathfrak{S}}((I \cup \{M\}), Q)$, corresponds to the set of tuples of constants that belong to the evaluation of Q over $K_{\mathbf{T}}$, for each GDE KB-solution J for I and $\mathcal{M}$ and $K \in \mathsf{Mod}((J \cup \{\mathcal{M}\}), \Sigma_t)$. Finally, generating $\mathsf{certain}_{\mathfrak{S}}((I \cup \{M\}), Q)$ is in Logspace.

5 Future Work

An interesting extension for this work would be defining a GDE setting with a target that contains egds and tgds constraints. Also, investigating GDE in a peer-to-peer setting might add interesting challenges to the problem.

References

1. Lawrence, M., Pottinger, R., Staub-French, S.: Data Coordination: Supporting Contingent Updates. In: VLDB (2011)
2. Philip A. Bernstein , Fausto Giunchiglia , Anastasios Kementsietsidis , John Mylopoulos , Luciano Serafini , llya Zaihrayeu : Data Management for Peer-to-Peer Computing: A Vision. pp. 89–94 (2002)
3. Arenas, M., Barceló, P. , Libkin, L.,Murlak, F.: Relational and XML Data Exchange. *Morgan & Claypool Publishers*, (2010).
4. Arenas, M. , Perez, J. ,Reutter, J.: Data exchange beyond complete data. In : PODS, pp. 83–94 (2011)
5. Fagin,R., Kolaitis, P. G. , Miller,R. J. , Popa, L.: Data exchange: semantics and query answering. In: Theoretical Computer Science, pp. 89–124 (2005) 31(4), pp. 761–791 (1984).
6. Kementsietsidis, A., Arenas, M., Miller, R. J. : Mapping data in peer-to-peer systems: Semantics and algorithmic issues. In: SIGMOD, pp. 325–336 (2003).