Foundations of Databases

Exercises – Datalog,
Datalog Evaluation

March 8, 2005

1. Refer to the Metro database. Give a datalog program which yields, for each pair of stations (a, b), the station c such that c is reachable (1) from both a and b; and (2) from either a or b.

   Test the program on the example database instance using DLV (`http://www.dlvsystem.com`)

2. Suppose we have a schema $\mathcal{S} = \{E1, E2\}$ of binary relations E_1 and E_2. Then, each database instance $\mathbf{I}$ can be viewed as representing two (finite) graphs $G_1(\mathbf{I}) = (adom(\mathbf{I}), E1(\mathbf{I}))$ and $G_2(\mathbf{I}) = (adom(\mathbf{I}), E2(\mathbf{I}))$ over the same set of nodes given by the active domain of $\mathbf{I}$.

   Write a datalog program which computes the set of pairs (a, b) such that there exists a (directed) path from node a to node b where edges from G_1 and G_2 alternate.

   Test the program using DLV on some instances.

3. Show that for any datalog program P and database instance $\mathbf{I}$, it holds that $adom(P(\mathbf{I})) \subseteq adom(P, \mathbf{I})$.

4. Consider a directed binary graph represented by a binary relation E as in Ex. (2). Give a datalog program which computes a binary relation T containing all pairs (a, b) such that there is a path of odd length from a to b.

   Test the program using DLV on some instances.

5. Prove the following statements: For every datalog program P,

(a) the operator $\mathbf{T}_P$ is monotonic, i.e., $\mathbf{K} \subseteq \mathbf{K}'$ implies $\mathbf{T}_P(\mathbf{K}) \subseteq \mathbf{T}_P(\mathbf{K}')$;

(b) $\mathbf{K} \in inst(sch(P))$ is a model of $\Sigma_P \Leftrightarrow \mathbf{T}_P(\mathbf{K}) \subseteq \mathbf{K}$;

(c) If $\mathbf{T}_P(\mathbf{K}) = \mathbf{K}$ then $\mathbf{K}$ is a model of Σ_P.

6. Show that every continuous operator $T : U \to U$ on a complete lattice $(U, \leq)$ is monotonic.

7. (Logical Consequence, for computational logicians) Consider the logical theory

$$\Sigma_{P,\mathbf{I}} = \Sigma_P \cup \{R(\vec{t}) \mid R \in edb(P), \vec{t} \in \mathbf{I}(R)\}$$

for the datalog program P and the database instance $\mathbf{I}$. Assume that in $\Sigma_{P,\mathbf{I}}$ at least one constant occurs. Show that $R(\vec{t}) \in P(\mathbf{I})$ holds iff $\Sigma_{P,\mathbf{I}} \models R(\vec{t})$ in classical logic, where the signature has as predicates the relation names in $sch(P)$ and has as constant symbols the elements of $adom(P, \mathbf{I})$. Show that the same holds if the signature has as constant symbols the elements of dom instead.

8. Show that datalog programs are monotonic, i.e., $P(\mathbf{I}) \subseteq P(\mathbf{J})$ whenever $\mathbf{I} \subseteq \mathbf{J}$.

9. Show that every expression of SPC Algebra in which elementary selections involve only equality atoms can be expressed in datalog, i.e., for each expression E in (unnamed) SPC Algebra on schema $\mathbf{R}$, there is a datalog program P_E with a designated relation $T \in idb(P_E)$ and $edb(P_E) = \mathbf{R}$ such that for every $\mathbf{I} \in inst(\mathbf{R})$, it holds that $P_E(\mathbf{I}) = E(\mathbf{I})$.

10. Apply Semi-Naive Evaluation to the following program:

$$\begin{aligned}
sgc(x,x) &\leftarrow person(x)\\
sgc(x,y) &\leftarrow par(x,x_1), sgc(x_1,y_1), par(y,y_1)\\
anc(x,y) &\leftarrow par(x,y)\\
anc(x,y) &\leftarrow anc(x,z), anc(z,y)\\
sgc_anc(x,y,z) &\leftarrow sgc(x,y), anc(x,z), anc(y,z)\\
sgc_anc(x,y,z) &\leftarrow sgc_anc(x,y,w), par(w,z)
\end{aligned}$$

on the instance $\mathbf{I}$ as follows:

$$\begin{aligned}
\mathbf{I}(person) &= \{ \langle ann\rangle, \langle bertrand\rangle, \langle charles\rangle, \langle dorothy\rangle,\\
&\quad\; \langle evelyn\rangle, \langle fred\rangle, \langle george\rangle, \langle hilary\rangle\}\\
\mathbf{I}(par) &= \{ \langle dorothy, george\rangle, \langle evelyn, george\rangle, \langle bertrand, dorothy\rangle,\\
&\quad\; \langle ann, dorothy\rangle, \langle ann, hilary\rangle, \langle charles, evelyn\rangle\}.
\end{aligned}$$

11. Consider Algorithm Semi-Naive.

    (a) Show that the algorithm terminates on each input $\mathbf{I}$ and that it is correct (assuming that P is a linearized program. (Hint: Use fixpoint semantics and $\mathbf{T}_P$ operator).

    (b) Suppose we replace in the algorithm $E_i[R_1, \ldots, D_i, \ldots, R_n]$ with $E_i[R_1, \ldots, D_{j_i}, \ldots, R_n]$, where $j_i \in \{1, \ldots, n\}$, for each i. Is the resulting more general algorithm correct? (Semi-Naive is the special case where $j_i = i$, for each i.)

12. Consider the adorned rule

$$R^{bf}(x, y) \leftarrow S^{bf}(x, y), T^{bf}(y, z), U^{ff}(u, v), W^{bf}(v, w).$$

    Explain why it makes sense to consider the second occurrence of v as bound.

13. Consider the rule

$$R(x, y, z) \leftarrow S(y, z), T(z, x).$$

    Construct adorned versions of this rule for R^{ffb} and R^{ffb}.