Constraint Extraction from SQL-queries

Toon Calders, Bart Goethals, and Adriana Prado

University of Antwerp, Belgium
{toon.calders, bart.goethals, adriana.prado}@ua.ac.be

1 Technical details

For extracting the constraints of a given SQL-mining query, we annotate every node of a corresponding relational algebra tree with a three-tuple $(\mathcal{A}, \mathcal{V}, \mathcal{O})$. In this three-tuple, $\mathcal{A}$ is the set of attributes $\{A_1, \ldots, A_n\}$ of that node, $\mathcal{V}$ is the set of *virtual mining views* [1] with the constraints for n $\{V_1[\phi_1], V_2[\phi_2], \ldots, V_m[\phi_m]\}$, and $\mathcal{O}$ a set of pairs $A_i \longrightarrow V_j$ denoting that the values in attribute A_i originate from the view V_j. Notice that the values in an attribute can originate from more than one view at the same time, namely if two views have been joined on this attribute.

The following tables contain the full technical details of how to compute the annotation of a node, given the annotations of its child(ren) and the operator represented by the node. We consider all operations from the relational algebra; i.e., π, σ, $\times$, $-$, $\cup$. The operation $\cup$ is handled by pushing it upwards in the expression tree; that is, a query involving $\cup$ is rewritten as $q_1 \cup \ldots \cup q_k$, where all q_i, $i = 1 \ldots k$ do not involve $\cup$. The algorithm is then applied to all q_i's in isolation, and in the end, the disjunction of the constraints is taken.

Furthermore, any selection $\sigma_{R_1.rid=R_2.rid}$ is replaced by the following selections: $\sigma_{R_1.rid=R_2.rid}\sigma_{R_1.sida=R_2.sida}\sigma_{R_1.sidc=R_2.sidc}\sigma_{R_1.sid=R_2.sid}$.

Leaf node	Annotation
Relation R (not a virtual mining view)	$(\{\}, \{\}, \{\})$
$SetsS$	$(\{S.sid, S.item\}, S[\,], \{S.sid \rightarrow S, S.item \rightarrow S\})$
$SupportsS$	$(\{S.sid, S.supp\}, S[\,], \{S.sid \rightarrow S, S.supp \rightarrow S\})$
$RulesR$	$(\{R.rid, R.conf, R.sida, R.sidc, R.sid\},$ $\{R[\,], S_a^R[\,], S_c^R[\,], S^R[\,],\},$ $\{R.rid \rightarrow R, R.conf \rightarrow R, R.sid \rightarrow S^R,$ $R.sida \rightarrow S_a^R, R.sidc \rightarrow S_c^R\})$

Operator	Child(ren) ann.	Annotation
$\pi_{A_1,\ldots,A_k}$	$(\mathcal{A}, \mathcal{V}, \mathcal{O})$	$(\{A_1, \ldots, A_k\}, \mathcal{V}, \{A \rightarrow V \in \mathcal{O} \mid A = A_i, 1 \leq i \leq k\})$
$\times$	$(\mathcal{A}_1, \mathcal{V}_1, \mathcal{O}_1)$, $(\mathcal{A}_2, \mathcal{V}_2, \mathcal{O}_2)$	$(\mathcal{A}_1 \cup \mathcal{A}_2, \mathcal{V}_1 \cup \mathcal{V}_2, \mathcal{O}_1 \cup \mathcal{O}_2)$
$-$	$(\mathcal{A}, \mathcal{V}_1, \mathcal{O}_1)$, $(\mathcal{A}, \mathcal{V}_2, \mathcal{O}_2)$	$(\mathcal{A}, \mathcal{V}_1 \cup \mathcal{V}_2, \mathcal{O}_1)$
$\sigma_{supp\theta c}$	$(\mathcal{A}, \mathcal{V}, \mathcal{O})$	$(\mathcal{A}, \mathcal{V}', \mathcal{O})$ $\mathcal{V}' = (\mathcal{V} - \mathcal{V}_{supp}) \cup \{V[\phi \wedge supp\theta c] \mid V[\phi] \in \mathcal{V}_{supp}\}$
$\sigma_{conf\theta c}$	$(\mathcal{A}, \mathcal{V}, \mathcal{O})$	$(\mathcal{A}, \mathcal{V}', \mathcal{O})$ $\mathcal{V}' = (\mathcal{V} - \mathcal{V}_{conf}) \cup \{V[\phi \wedge conf\theta c] \mid V[\phi] \in \mathcal{V}_{conf}\}$
$\sigma_{item=i}$	$(\mathcal{A}, \mathcal{V}, \mathcal{O})$	$(\mathcal{A}, \mathcal{V}', \mathcal{O})$ $\mathcal{V}' = (\mathcal{V} - \mathcal{V}_{item}) \cup \{V[\phi \wedge (i \in I)]) \mid V[\phi] \in \mathcal{V}_{item}\}$
$\sigma_{sid_1=sid_2}$	$(\mathcal{A}, \mathcal{V}, \mathcal{O})$	$(\mathcal{A}, \mathcal{V}', \mathcal{O}')$ Let $\Phi = \bigwedge_{S[\phi] \in \mathcal{V}_{sid_1} \cup \mathcal{V}_{sid_2}} \phi.$ $\mathcal{V}' = (\mathcal{V} - (\mathcal{V}_{sid_1} \cup \mathcal{V}_{sid_2})) \cup \{V[\Phi] \mid \exists \psi : V[\psi] \in \mathcal{V}_{sid_1} \cup \mathcal{V}_{sid_2}\}$ $\mathcal{O}' = \mathcal{O} \cup \{A \rightarrow S_1 \mid sid_1 \rightarrow S_1, \exists S_2 : sid_2 \rightarrow S_2, A \rightarrow S_2\}$
$\sigma_{rid_1=rid_2}$	$(\mathcal{A}, \mathcal{V}, \mathcal{O})$	$(\mathcal{A}, \mathcal{V}', \mathcal{O}')$ Let $\Phi = \bigwedge_{R[\phi] \in \mathcal{V}_{rid_1} \cup \mathcal{V}_{rid_2}} \phi.$ $\mathcal{V}' = (\mathcal{V} - (\mathcal{V}_{sid_1} \cup \mathcal{V}_{sid_2})) \cup \{R[\Phi] \mid \exists \psi : R[\psi] \in \mathcal{V}_{rid_1} \cup \mathcal{V}_{rid_2}\}$ $\mathcal{O}' = \mathcal{O} \cup \{A \rightarrow R_1 \mid rid_1 \rightarrow R_1, \exists R_2 : rid_2 \rightarrow R_2, A \rightarrow R_2\}$
Any other selection	$(\mathcal{A}, \mathcal{V}, \mathcal{O})$	$(\mathcal{A}, \mathcal{V}, \mathcal{O})$

θ is one of $=, \leq, \geq, <, >$, c is a constant number
sid_1 and sid_2 are set identifiers.
rid_1 and rid_2 are rule identifiers.
$\mathcal{V}_A = \{V[\phi] \in \mathcal{V} \mid A \rightarrow V \in \mathcal{O}\}$.

References

[1] T. Calders, B. Goethals and A. Prado. Integrating Pattern Mining in Relational Databases. In *10th PKDD Conference* (2006) 454–461.