Provenance of aggregate queries with HAVING clause in ProvSQL.

Pratik Karmakar Aryak Sen Pierre Senellart

R is a relation on set of attributes U.
We consider $U^{GB} \subseteq U$ and $U^{AGG} \subseteq U$ and $U^{GB} \cap U^{AGG} = \phi$.
For each tuple t,

$$T = \{t^* \in \mathrm{Supp}(R) | \forall u \in U^{GB}, t(u) = t^*(u)\}.$$

Extending on the semantics of aggregate GROUP BY queries [1] we express the provenance of HAVING queries as:

$$q := \sigma_{SUM} = c$$

$$\mathrm{Provenance}(q) = \delta(\sum_{t_i \in T} t_i) * [\sum K \otimes SUM_{t_i \in T} t_i \otimes c_i = c \otimes \mathbb{1}]$$

1 Formula Semiring

$$\mathcal{K}_{formula} = (K, \oplus, \otimes, \mathbb{0}, \mathbb{1}, \delta, \ominus)$$

$K \leftarrow$ Set of strings

$$\oplus(k_1, \dots, k_n) = \begin{cases} 0_k & \text{if } n = 0 \\ k_1 & \text{if } n = 1 \\ k_1 \oplus (k_2, \dots, k_n) & \text{if } n > 1 \end{cases}$$

$$\otimes : K^n \to K$$

$$\ominus : K \times K \to K : k_1 \ominus k_2 = (+k_1 + \ominus + k_2 +)$$

$$\delta : K \to K$$

$$\delta(k) = \begin{cases} \delta(+k+) & \text{if } k[0] \neq' C' \\ \delta + k & \text{if } k[0] =' C' \end{cases}$$

$$Cmp : K^2 \to K$$
$$op \in \{=, \neq, <, \leq, >, \geq\}$$
$$Cmp(k_1, op, k_2) = [+k_1 + op + k_2+]$$

$$agg(\gamma, \{q_i\}) : \{SUM, MIN, MAX, PROD\} \times K^n \to K$$
$$SUM : \sum q_i$$
$$MIN : min(k_1, \dots, k_n)$$
$$MAX : max(k_1, \dots, k_n)$$
$$PROD : \prod k_i$$

U be the set of attributes on domain D.
Tuples are $tup(U) = \{t : U \to D\}$ A K-relation is $R : tup(U) \to K$

- Empty relation: $[Q](t) = 0_K$
- SELECTION: $[\sigma_\theta(R)](t) = \delta(Cmp(\theta_1)) \otimes \delta(Cmp(\theta_2)) \otimes \cdots \otimes R(t)$
- NATURAL JOIN: $[R \bowtie S](t) = R(t_R) \otimes S(t_S)$, given $Q = R \bowtie S \forall t$ with projection t_R, t_S
- RENAME: $\rho_{A \to B}(R), [\rho_{A \to B}(R)](t) = R(t[A \mapsto B])$
- DIFF: $[R - S](t) = R(t) \ominus S(t)$
- PROJECTION: $U \subseteq Schema(R), u \in tup(U)$
  $[\Pi_U(R)](u) = \oplus_{t[U]=u} R(t)$
- UNION: $[R \cup S](t) = R(t) \oplus S(t)$

Let G be the set of attributes in a GROUP BY clause,

$$G \subseteq Schema(R)$$

and aggregate function $f \in \{SUM, PROD, MIN, MAX\}$ over attributes $A \in U \setminus G$.
And,

$$\delta = \gamma_{G,f(A)}(R),$$
$$\delta(g) = agg(\gamma, \{R(t)|t[G] = g\}) \forall group \in tup(G)$$

1.1 HAVING only for constant C to be compared against f

$$\delta = \gamma_{G,f(A)}(R) : g \mapsto \delta(g) = f(\{R(t)|t[G] = g\}) \in K$$

Only one provenance token $\delta(g)$ for each group g in the GROUP BY clause. We give semantics for $f(A) opC$:

$$[\gamma_{G,f(A)}(R) \text{ HAVING } f(A)opC] = (g \mapsto \gamma([\delta(g)opC]) \otimes \delta(g))$$
$$\text{where } op \in \{=, \neq, <, \leq, >, \geq\}$$

[1] [2]

References

[1] Amsterdamer, Y., Deutch, D., Tannen, V.: Provenance for aggregate queries. In: Proceedings of the thirtieth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems. pp. 153–164 (2011)

[1] In deterministic scenario, provenance of HAVING queries is just Boolean existence onus the comparison operator??

[2] For probabilistic databases, provenance of HAVING queries is to be computed using the DP algo.