Introduction to Databases

Relational Algebra

under bag semantics

Dr Paolo Guagliardo

(v26.0)

Problem 1. Under set semantics, the natural join of two expressions with the same sets of attributes is equivalent to their intersection, that is, $R \bowtie S \equiv R \cap S$ whenever R and S are over the same attributes. Under bag semantics, natural join is defined in precisely the same way as under sets (but multiplicities are taken into account, of course), but the above equivalence no longer holds. Your task here is to exhibit a database that witnesses this fact.

Solution. It is enough to consider relations R and S over a single attribute, say A, and the database instance where R consists of one occurrence of some constant a and S consists of *two* occurrences of the same constant a. Then, $R \cap S$ will produce a table with a single attribute A, consisting of one occurrences of a. On the other hand, $R \bowtie S$ amounts to

$$\pi_A\Big(\sigma_{A=A'}\big(R \times \rho_{A \to A'}(S)\big)\Big)$$

Now, the innermost Cartesian product produces the following table:

A	A'
a	a
a	a

As both of the above records satisfy the selection condition $A = A'$, the answer to the overall query is a table with a single attribute A, consisting of two occurrences of a (rather than one). The existence of a database on which the two expressions produce different answers *proves* that they are not equivalent.

Problem 2. For relations R and S with disjoint sets of attributes, the (left) *semi-join operation* $R \ltimes_\theta S$ in Relational Algebra under bag semantics is defined as follows:

$$\bar{r} \in_k R \ltimes_\theta S \iff \bar{r} \in_k R \text{ and there is } \bar{s} \in S \text{ such that } \bar{r}\bar{s} \text{ satisfies } \theta$$

Intuitively, $R \ltimes_\theta S$ returns all records of R, with the multiplicity with which they occur in R, for which there is a record in S such that their concatenation satisfies the condition θ.

Write a Relational Algebra expression that, under bag semantics, is equivalent to the semi-join operation defined above, and uses only primitive operations (possibly including duplicate elimination).

Solution. Let X denote the set of attributes of R. The expression that defines the left semi-join under set semantics, namely $\pi_X\big(\sigma_\theta(R \times S)\big)$, does not produce the desired result under bags semantics. Indeed, a record $\bar{r} \in_k R$, for which there is at least one match in S, will occur $k \cdot m$ times in the output, where is m is the sum of the multiplicities of all suitable records in S.

Even though it is not exactly what we need, the above expression is still a very good starting point. The problem with it is that it returns qualifying records of R with a multiplicity that is at least the multiplicity with which they occur in R, while we need them with the same multiplicity they have in R. To achieve this, all we have to do is take the intersection with R, and rewrite it in terms of difference to get an expression that uses only primitive operations:

$$R \ltimes_\theta S \equiv R \cap \pi_X\big(\sigma_\theta(R \times S)\big) \equiv R - \big(R - \pi_X\big(\sigma_\theta(R \times S)\big)\big)$$