Lecture 19

Tuesday, 12 November 2019 14:32

Probabilistic database: distribution over database instances

Possible worlds (support), instances: $w = \{I_1, \dots, I_n\}$
$p: w \mapsto [0,1]$
$\sum_{i=1}^{n} p(I_I) = 1$

Marginal distribution of tuple t
$\Pr(t) = \sum_{i:t \in I_i} p(I_i)$

Independence of tuples
$\Pr(t_1 \wedge t_2) = \Pr(t_1) \Pr(t_2)$

Query evaluation
$q(w,p) = \{(q(I_1), p(I_1)), \dots, (q(I_n), p(I_n))\}$
$\Pr(t \in q(w,p))$

Tuple-independent: each tuple is associated with a single probability (does not support correlations between tuples)

Block-independent
- A block contains tuples with the same key value
- Sum of tuple probabilities in each block is less than 1
- In every block, either zero or one tuple is true (disjoint, to satisfy the key constraint)
- Probabilities between blocks are independent

Tuple-independent model is block-independent with all attributes being the key

Modify relational operators to handle probabilities
Selection: simple filtering, do not change probabilities
Independent join: product of a pair of joined tuple is the product of the tuples
Independent projection: $1 - \prod_{i=1}^{a}(1 - \Pr(A = i))$

Non-independent join: need to first project the join variables to obtain a tuple-independent relation, then multiply each probability as a join (safe plan)

Example of query that does not admit a safe plan:
$q(\square): -R(x,z), S(x,w), T(w,z)$