1. Motivation

Goal: manage uncertain information in different application scenarios: data integration, scientific data collections, census ...

Features of MayBMS
- scalable DBMS for supporting uncertain and probabilistic data
- purely relational representation of attribute-level uncertainty
- efficient query processing
- query language for probabilistic databases

2. U-relational Databases

R[SSN,N,MS]: personal information

<table>
<thead>
<tr>
<th>TID</th>
<th>SSN</th>
<th>V₁ → D₁</th>
<th>U₁₁₁</th>
</tr>
</thead>
<tbody>
<tr>
<td>t₁</td>
<td>185</td>
<td>x → 1</td>
<td></td>
</tr>
<tr>
<td>t₂</td>
<td>785</td>
<td>x → 2</td>
<td></td>
</tr>
<tr>
<td>t₃</td>
<td>185</td>
<td>y → 1</td>
<td></td>
</tr>
<tr>
<td>t₄</td>
<td>186</td>
<td>y → 2</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>TID</th>
<th>MS</th>
<th>V₁ → D₁</th>
<th>U₁₁₁</th>
</tr>
</thead>
<tbody>
<tr>
<td>t₁</td>
<td>1</td>
<td>z → 1</td>
<td></td>
</tr>
<tr>
<td>t₂</td>
<td>2</td>
<td>z → 2</td>
<td></td>
</tr>
<tr>
<td>t₃</td>
<td>1</td>
<td>w → 1</td>
<td></td>
</tr>
<tr>
<td>t₄</td>
<td>2</td>
<td>w → 2</td>
<td></td>
</tr>
<tr>
<td>t₅</td>
<td>3</td>
<td>w → 3</td>
<td></td>
</tr>
<tr>
<td>t₆</td>
<td>4</td>
<td>w → 4</td>
<td></td>
</tr>
</tbody>
</table>

3. Query Language

World-set Algebra
- extend relational algebra with uncertainty-specific constructs e.g.
  - conf: confidence computation
  - repair-by-key: create the possible repairs of an instance violating a key constraint
  - semantics: evaluate the query in each world
- properties
  - generality: independent from representation details
  - conservative over relational algebra: right degree of expressive power
  - efficient evaluation: simple encoding of positive relational algebra + possible into positive relational algebra queries on U-relational databases

4. Query Evaluation

- possible(π₁₁₁(U₁₁₁ ∧ U₁₁₁))
- merge(π₁₁₁(U₁₁₁), π₁₁₁(U₁₁₁))
- Query on U-relational databases:
  \[ π₁₁₁(U₁₁₁ ∧ U₁₁₁) \]

- repair-key_{SSN}(R)

- Query on column-stores
  \[ π_{merge}(π_{merge}(π_{merge}(\pi_{join}(\pi_{SSN}R, \pi_{SSN}(\pi_{SSN}S)))) \]

- push merge up
  \[ T \]

5. Confidence Computation

conf of (A:1) = probability of the world-set defined by

\[ \{(x→1), (x→2,y→1), (x→2,z→1), (u→1,v→1), (u→2)\} \]

6. Experiments

- extended TPC-H population generator 2.6 to generate U-relational databases
- parameters: scale (s), uncertainty ratio (x), correlation ratio (z), max alternatives per field (m), drop after correlation (p)
- each generated world has the sizes of relations and join selectivities of the original TPC-H one-world case
- queries translated into SQL and run on PostgreSQL

- Fig. 1: Number of worlds and size in MB of the U-relational db for different scale, uncertainty and correlation ratios.
- Fig. 2: Query evaluation
- Fig. 3: Attribute- vs. tuple-level representation
- Fig. 4: TPC-H (scale=10^10) world sizes
- Fig. 5: TPC-H (scale=10^10) world sizes in MB

7. Selected Publications
1. Fast and simple relational processing of uncertain data, ICDE’08
2. Conditioned Probabilistic Databases, Technical report ’08
3. MayBMS: Managing incomplete information with probabilistic world-set decompositions, ICDE’07, Demonstration
4. 10^10 worlds and beyond: efficient representation and processing of incomplete information, ICDE’07
5. From complete to incomplete information and back, SSYMID’07