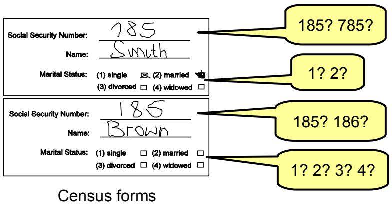
MayBMS: A System for Managing Large Uncertain and Probabilistic Databases

Cornell University Computer Science

Lyublena Antova, Cornell University
Christoph Koch, Cornell University
Dan Olteanu, Oxford University

1. Motivation

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



Census forms

There are 2 * 2 * 2 * 4 = 32 possible instances of the forms information!

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

$U_{R[SSN]}$	TID	SSN	$V_1 \to D_1$
	t_1	185	$x \to 1$
	t_1	785	$x \to 2$
	t_2	185	$y \to 1$
	t_2	186	$y \to 2$

$U_{R[MS]}$	TID	MS	$V_1 \to D_1$
	t_1	1	$z \to 1$
	t_1	2	$z \to 2$
	t_2	1	$w \to 1$
	t_2	2	$w \to 2$
	t_2	3	$w \to 3$
	t_2	4	$w \to 4$

$U_{R[N]}$	TID	N	$V_1 \to D_1$
	t_1	Smith	
	t_2	Brown	

world table: prob. distribution of the variables

W	V→D	Pr
	$x \to 1$	0.4
	$x \to 2$	0.6
	$y \to 1$	0.7
	$y \to 2$	0.3
	$z \to 1$	0.8
	$z \to 2$	0.2
	$w \to 1$	0.25
	$w \to 2$	0.25
	$w \to 3$	0.25
	$w \to 4$	0.25

S[SSN,ST]: credit status

$U_{S[SSN,ST]}$	TID	SSN	ST	$V_1 \to D_1$
	s_1	185	bad	$w \to 3$
	s_1	185	good	$w \to 4$

encode attribute alternatives and correlations with variables

Construct a possible world: pick a value for each variable

R	SSN	Name	MS
t_1	785	Smith	2
t_2	186	Brown	4

S	SSN	ST
s_1	185	good

Probability of the world: 0.6*0.3*0.2*0.25=0.009

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
  - **assert**: remove worlds violating a constraint
- ✓semantics: evaluate the query in each world
- ✓properties
  - generic: 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

a) possible($\pi_N(\sigma_{MS=3}(R))$)

Query on U-relational databases:

$\pi_N(\sigma_{MS=3}(U_{R[SSN]} \bowtie_{\varphi\&\psi} U_{R[MS]}))$

merge($\pi_N R$, $\pi_{MS} R$)

φ=l.TID=r.TID

ψ=($l.V_1$=$r.V_1 \to l.D_1$=$r.D_1$) & ($l.V_2$=$r.V_1 \to l.D_2$=$r.D_1$) ← ensure consistent variable assignment

b) repair-key$_{SSN}$ (R)

new variable for each non-unique SSN value

$U_{R[SSN]}$	TID	SSN	$V_1 \to D_1$	$V_2 \to D_2$
	t_1	185	$x \to 1$	$x_1 \to 1$
	t_1	785	$x \to 2$	
	t_2	185	$y \to 1$	$x_1 \to 2$
	t_2	186	$y \to 2$	

c) $\pi_{MS}(R \bowtie_{SSN} (\sigma_{ST=bad} S))$

Query on column-stores

π_{MS}(**merge**($\pi_{SSN}R$, $\pi_{MS}R$) $\bowtie_{SSN} \sigma_{ST=bad}(S)$) ≡

π_{MS}(**merge**($\pi_{\{\}}(\pi_{SSN} R \bowtie_{SSN} \sigma_{ST=bad}(S))$, $\pi_{MS}R$))

T

push **merge** up

intermediate result:

$U_{T[]}$	TID	$V_1 \to D_1$	$V_2 \to D_2$
	t_1,s_1	$x \to 1$	$w \to 3$
	t_2,s_1	$y \to 1$	$w \to 3$

final result:

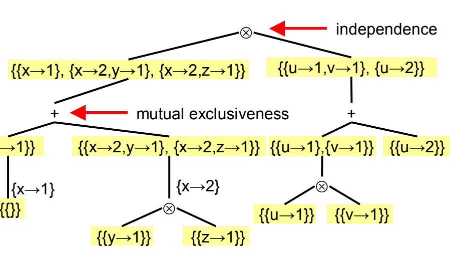
$U_{P[MS]}$	TID	MS	$V_1 \to D_1$	$V_2 \to D_2$	$V_3 \to D_3$
	t_1,s_1	1	$x \to 1$	$w \to 3$	$z \to 1$
	t_1,s_1	2	$x \to 2$	$w \to 3$	$z \to 2$
	t_2,s1	3	$y \to 1$	$w \to 3$	

5. Confidence Computation

$U_{R[A]}$	A	$V_1 \to D_1$	$V_2 \to D_2$
	1	$x \to 1$	
	1	$x \to 2$	$y \to 1$
	1	$x \to 2$	$z \to 1$
	1	$u \to 1$	$v \to 1$
	1	$u \to 2$	

confidence 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 PostgresSQL

s	z	TPC-H dbsize	#worlds	dbsize	#worlds	dbsize
0.5	0.1	853	$10^{43368.0}$	3843	$10^{3.97e+06}$	5427
0.5	0.5	853	$10^{25528.9}$	3856	$10^{2.33e+06}$	6682
1	0.1	1706	$10^{87203.0}$	7683	$10^{7.94e+06}$	11264
1	0.5	1706	$10^{51290.9}$	7712	$10^{4.66e+06}$	13312
		X=0.0	X=0.001		X=0.1	

Fig. 1: Number of worlds and size in MB of the U-relational db for different scale, uncertainty and correlation ratios.

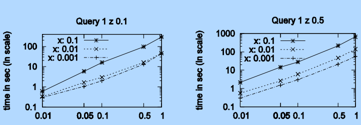


Fig. 2: Query evaluation

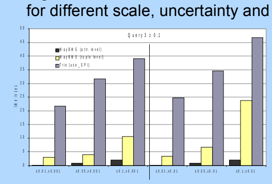


Fig. 3: Attribute- vs. tuple-level representation

7. Selected Publications

1. Fast and simple relational processing of uncertain data, ICDE'08
2. Conditioning Probabilistic Databases, Technical report '08
3. MayBMS: Managing incomplete information with probabilistic world-set decompositions, ICDE'07, Demo paper
4. 10^10^6 worlds and beyond: efficient representation and processing of incomplete information, ICDE'07
5. From complete to incomplete information and back, SIGMOD'07