Preservation Properties and Databases

Anuj Dawar
University of Cambridge

joint work with Albert Atserias and Phokion Kolaitis

Imperial College, 24 November 2003

Relational Databases

$$\mathcal{C}inema = \{Movies[3], Location[3], Guide[3]\}$$

Movies	Title	Director	Actor
	Magnolia	Anderson	Moore
	Magnolia	Anderson	Cruise
	Spiderman	Raimi	Maguire
	Spiderman	Raimi	Dunst
	...		
	Rocky	Avildsen	Stallone
	RockyII	Stalone	Stallone

Guide	Title	Cinema	Time
	Rocky	Warner	12:00
	Spiderman	Picturehouse	19:00
	...		
	Spiderman	Phoenix	19:00
	Magnolia	Picturehouse	22:00

Location	Cinema	Address	Tel
	Picturehouse	Cambridge	504444
	Phoenix	Oxford	512526
	Warner	Cambridge	560225

Relational Algebra

In relational algebra, queries are built up from

Base relations:	R
Singleton constant relations:	$\{\langle a\rangle\}$

using

select:	$\sigma_{j=a}(q)$ or $\sigma_{j=k}(q)$
project:	$\pi_{j_1,\ldots,j_k}(q)$
join:	$q_1 \bowtie q_2$
union:	$q_1 \cup q_2$
difference:	$q_1 - q_2$

SPJU Algebra

All the operators of relational algebra other than difference are *monotone*:

> Adding new facts to the database cannot remove a tuple from the result.

Most queries actually used are written using only SPJU (*select-project-join-union*).

The answers do not rely on a *closed-world assumption*.

Does allowing *difference* allow us to express any new monotone queries?

Relational Calculus

Codd in 1972 introduced the relational calculus (based on first-order logic) and equivalent to the relational algebra.

Conjunctive Queries:

$$q(x,y) \leftarrow \mathit{Movies}(z_1, \text{“Almodovar”}, z_2), \mathit{Guide}(x, z_1, z_3), \mathit{Location}(x, y, z_4)$$

expresses the query

$$\{x, y \mid \exists z_1, \ldots, z_4\ \mathit{Movies}(z_1, \text{“Almodovar”}, z_2) \land \mathit{Guide}(x, z_1, z_3) \land \mathit{Location}(x, y, z_4)\}$$

Disjunction is expressed by *multiple rules*.

Existential Positive Logic

Adding negation and universal quantification gives us the full-power of relational algebra.

The existential-positive fragment is exactly equivalent to the SPJU algebra. (also known as *recursion-free Datalog*).

The closed-world assumption is even more important to the semantics of negation, as answers may by *infinite*.

Issues of *safety* and *domain independence*.

Assume finite domain?

Homomorphism

A first-order query is interpreted over some structure (or database state):

$$\mathbb{A} = (A, R_1^{\mathbb{A}}, \ldots, R_m^{\mathbb{A}})$$

where A is the *domain* and the $R_i^{\mathbb{A}}$ are relations over A, interpreting the symbol R_i.

Given $\mathbb{A} = (A, R_1^{\mathbb{A}}, \ldots, R_m^{\mathbb{A}})$ and $\mathbb{B} = (B, R_1^{\mathbb{B}}, \ldots, R_m^{\mathbb{B}})$, a function $f : A \to B$ is a *homomorphism* if:

$$R_i^{\mathbb{A}}(\mathbf{a}) \quad \text{implies} \quad R_i^{\mathbb{B}}(f(\mathbf{a}))$$

Restrictions

A homomorphism $f : \mathbb{A} \rightarrow \mathbb{B}$ may or may not be *injective*.

It may or may not be *surjective*.

If it is injective and

$$R_i^{\mathbb{A}}(\mathbf{a}) \quad \text{if, and only if,} \quad R_i^{\mathbb{B}}(f(\mathbf{a}))$$

we say it is an *embedding*, or that $\mathbb{B}$ is an extension of $\mathbb{A}$.

q is *preserved* under homomorphisms if

$$\mathbf{a} \in q(\mathbb{A}) \quad \text{implies} \quad f(\mathbf{a}) \in q(\mathbb{B})$$

Preservation Properties

Classical theorems of model theory tell us that these various restrictions have syntactic counterparts:

If q is a query of the relational calculus then:

q is equivalent to an *existential-positive* query if, and only if, it is preserved under homomorphisms.

q is equivalent to a *positive* query if, and only if, it is preserved under surjective homomorphisms. **Lyndon**.

q is equivalent to an *existential* query if, and only if, it is preserved under embeddings. **Łoś-Tarski**

Finite Structures

Does this mean that the relational calculus is conservative over the SPJU algebra for monotone queries?

Unfortunately, the preservation results do not carry over if we assume that the *domain is finite*.

q is preserved under homomorphisms
implies
q is equivalent to an *existential-positive* query.

Restriction to finite structures weakens both hypothesis and consequent.

Preservation in the Finite

Both the Lyndon and Łoś-Tarski properties are known to fail when we restrict ourselves to finite structures.

The status of the homomorphism preservation property in the finite is a long-standing open question, with extensive literature.

Datalog

Datalog extends conjunctive rules with *recursion*

$$\begin{aligned} T(x,y) &\leftarrow E(x,y) \\ T(x,y) &\leftarrow E(x,z), T(z,y). \end{aligned}$$

defines the transitive closure T of a relation E.

This is not definable in the first-order logic.

Every query definable in **Datalog** is preserved under homomorphisms.

Datalog vs. First-Order

Ajtai and Gurevich (1994) showed that every query that is expressible in both Datalog and first-order logic is equivalent to an existential-positive query.

> This can be seen as a partial result towards the goal of showing the homomorphism preservation property.

Minimal Models

Say that a structure $\mathbb{A}$ is a *minimal model* of a query q if, for some $\mathbf{a} \in q(\mathbb{A})$ there is no proper submodel $\mathbb{B}$ of $\mathbb{A}$ with $\mathbf{a} \in q(\mathbb{B})$.

The collection of minimal models of q "generates", through homomorphisms, all models of q.

If q has finitely many minimal models, it can be expressed as an existential-positive query.

Ajtai-Gurevich

The Ajtai-Gurevich proof can be naturally broken into two parts:

1. the collection of minimal models of a first-order q which is closed under homomorphisms satisfy a certain combinatorial condition C; and
2. if the minimal models of a Datalog query satisfy C, then there are only finitely many of them.

We generalise this to many other interesting cases.

Combinatorial Condition

The combinatorial condition C:

Lemma (Ajtai-Gurevich)
For any first-order φ whose models are closed under homomorphisms and any positive integer s there are d and m such that if $\mathbb{A}$ is a *minimal* model of φ and $B \subset A$ has $|B| < s$ then

$\mathbb{A} - B$ does not contain a d-scattered set of m elements.

Tree-Width

The collection of minimal models of a Datalog query have bounded *tree-width*.

Tree-width is a measure of how a *tree-like* a structure is.

It is originally defined for graphs, though easily extends to other relational structures.

It has proved extremely useful in algorithm design and analysis, including algorithms for database query evaluation.

Tree-Width

For a graph $G = (V, E)$, a *tree decomposition* of G is a relation $D \subset V \times T$ with a tree T such that:

- for each $v \in V$, the set $\{t \mid (v, t) \in D\}$ forms a connected subtree of T; and
- for each edge $(u, v) \in E$, there is a $t \in T$ such that $(u, t), (v, t) \in D$.

The *tree-width* of G is the least k such that there is a tree T and a tree-decomposition $D \subset V \times T$ such that for each $t \in T$,

$$|\{v \in V \mid (v, t) \in D\}| \leq k + 1.$$

Examples

- Trees have tree-width 1.
- Cycles have tree-width 2.
- The clique K_k has tree-width $k-1$.
- The $m \times n$ grid has tree-width $\min(m,n)$.

For a general relational structure $\mathbb{A}$, we define its tree-width to be the tree-width of the graph $\Gamma\mathbb{A}$ in which $a \sim b$ if a and b occur in the same tuple of some relation.

Result and Consequences

We show that, if a collection of structures has property C, and has bounded-tree width, then there are only finitely many of them.

Consequences:

- the Ajtai-Gurevich result (also to infinitary extension).
- for any k, if $\mathcal{T}_k$ is the collection of structures of tree-width less than k, then the homomorphism preservation property holds on $\mathcal{T}_k$
- If q is a first-order query preserved under homomorphisms, q is equivalent to an existential-positive query if, and only if, the collection of its minimal models has bounded tree-width.

Graph Minors

We say that a graph $G = (V, E)$ is a minor of graph $H = (U, F)$, (written $G \prec H$) if there is a graph $H' = (U', F')$ with $U' \subseteq U$ and $F' \subseteq F$ and a surjective map

$$M : U' \to V$$

such that

- for each $v \in V$, $M^{-1}(v)$ is a connected subgraph of H'; and
- for each edge $(u, v) \in E$, there is an edge in F' between some $x \in M^{-1}(u)$ and some $y \in M^{-1}(v)$.

Graph Minors

Less formally, a a graph $G = (V, E)$ is a minor of graph $H = (U, F)$, (written $G \prec H$) if we can get G from H by a sequence of operations of:

- delete a node
- delete an edge
- contract an edge.

Facts about Graph Minors

G is planar if, and only if, $K_5 \not\prec G$ and $K_{3,3} \not\prec G$.

If $G \prec H$, then $\text{tree-width}(G) \leq \text{tree-width}(H)$.

The relation $\prec$ is transitive.

If $\text{tree-width}(G) < k - 1$, then $K_k \not\prec G$.

$K_k \prec K_{k-1,k-1}$.

A class of graphs $\mathcal{C}$ has bounded tree-width if, and only if, there is some grid G such that $G \not\prec H$ for any $H \in \mathcal{C}$.

Theorem (Robertson-Seymour)
In any infinite collection $\{G_i \mid i \in \omega\}$ of graphs, there are i, j with $G_i \prec G_j$.

Result and Consequences

We show that, if a collection $\mathcal{G}$ of graphs has property C, and for some graph H, H is not a minor of any graph in $\mathcal{G}$, then there are only finitely many graphs in $\mathcal{G}$.

Consequences:

- If $\mathcal{C}$ is a class of structures such that there is a graph H with $H \not\prec \Gamma\mathbb{A}$ for all $\mathbb{A} \in \mathcal{C}$ and $\mathcal{C}$ is closed under taking minors, then the homomorphism preservation property holds on $\mathcal{C}$.
- A first-order query q preserved under homomorphisms is equivalent to an existential-positive query if, and only if, the collection of its minimal models excludes some minor.