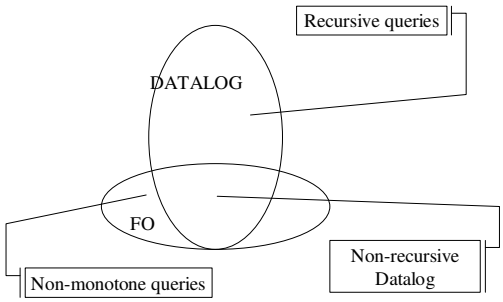
Conjunctive Queries

Thursday, February 8, 2001

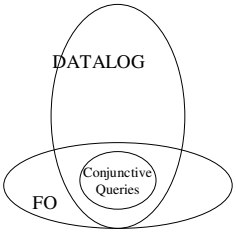
Outline

- Definition of conjunctive queries
- Query containment
- Query equivalence
- Query minimization

Picture So Far



Picture for This Lecture



Conjunctive Queries

- ***Definition*** A conjunctive query is exactly one Datalog rule

  or:

- ***Definition*** A conjunctive query is FO restricted to $R(x_1,..., x_n), \wedge, \exists$ (missing are $\vee, \neg, \forall$)

Examples

Employee(x), ManagedBy(x,y), Manager(y)

- Find all employees having the same manager as Smith:

  A(x) :- ManagedBy("Smith",y), ManagedBy(x,y)

Examples

Employee(x), ManagedBy(x,y), Manager(y)

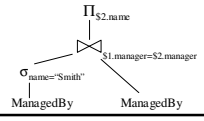
- Find all employees having the same director as Smith:

A(x) :- ManagedBy("Smith",y), ManagedBy(y,z),
ManagedBy(x,u), ManagedBy(u,z)

Equivalent Formulations

Relational Algebra:

- Conjunctive queries correspond precisely to σ_C, Π_A, x (missing: U, -)
- A(x) :- ManagedBy("Smith",y), ManagedBy(x,y)



Equivalent Formulations

SQL:

- Conjunctive queries correspond precisely to single select-from-where blocks

```
select m2.name
from ManagedBy m1, ManagedBy m2
where m1.name="Smith" AND
      m1.manager=m2.manager
```

Conjunctive Queries

- Most useful class of queries
- Also enjoys remarkable, positive properties
- Focus of research during 70's, 80's
- Still focus of research in the 00's
- We discuss the most celebrated property of conjunctive queries: containment is decidable

Query Containment

- ***Definition*** Given two queries q_1, q_2, we say that q_1 is contained in q_2 if for every database **D**, $q_1(\mathbf{D}) \subseteq q_2(\mathbf{D})$.
- Notation: $q_1 \subseteq q_2$
- Obviously: if $q_1 \subseteq q_2$ and $q_2 \subseteq q_1$ then $q_1 = q_2$.

Examples of Query Containments

$q_1(x,y)$:- R(x,u),R(v,u),R(v,y)

$q_2(x,y)$:- R(x,u),R(v,u),R(v,w),R(t,w),R(t,y)

Then $q_1 \subseteq q_2$ (why ?)

Examples of Query Containments

$q_1(x)$:- R(x,u), R(u,"Smith"), R(u,"Fred"), R(u, u)

$q_2(x)$:- R(x,u), R(u,v), R(u,"Smith"), R(w,u)

Then $q_1 \subseteq q_2$ (why ?)

Query Containment

- Recall: query containment for FO is undecidable
- ***Theorem*** Query containment for conjunctive queries is decidable and NP-complete.

Query Containment

The most interesting part: *how* we check $q_1 \subseteq q_2$

- The *canonical database* and the *canonical tuple* for q_1:
  - Canonical database: $\mathbf{D_{q1}} = (D, R_1, \ldots, R_k)$ where:
    - D = all variables and constants in q_1
    - $R_1, \ldots, R_k$ = the body of q_1
  - Canonical tuple: t_{q1} = the head of q_1

Examples of Canonical Databases

- $q_1(x,y)$:- R(x,u),R(v,u),R(v,y)
- $\mathbf{D}_{q1} = (D, R)$
  - D={x,y,u,v}
  - R =

x	y
v	u
v	y

- $t_{q1} = (x,y)$

Examples of Canonical Databases

$q_1(x)$:- R(x,u), R(u,"Smith"), R(u,"Fred"), R(u, u)

- $\mathbf{D}_{q1} = (D, R)$
  - D={x,u,"Smith","Fred"}
  - R =

x	u
u	"Smith"
u	"Fred"
u	u

- $t_{q1} = (x)$

Checking Containment

Theorem: $q_1 \subseteq q_2$ iff $t_{q1} \in q_1(\mathbf{D}_{q1})$.

Example:
$q_1(x,y)$:- R(x,u),R(v,u),R(v,y)
$q_2(x,y)$:- R(x,u),R(v,u),R(v,w),R(t,w),R(t,y)

- D={x,y,u,v}
- R =

x	y
v	u
v	y

$t_{q1} = (x,y)$

- Yes, $q_1 \subseteq q_2$

Query Homeomorphisms

- How do we *evaluate* q_2 on $\mathbf{D}_{q1}$?
- A ***homeomorphism*** $f : q_2 \rightarrow q_1$ is a function $f: var(q_2) \rightarrow var(q_1) \cup const(q_1)$ such that:
  - $f(body(q_2)) \subseteq body(q_1)$
  - $f(canonicalTuple(q_2)) = canonicalTuple(q_1)$

Example of Query Homeomorphism

$var(q_1) = \{x, u, v, y\}$

$var(q_2) = \{x, u, v, w, t, y\}$

$q_1(x,y)$:- R(x,u),R(v,u),R(v,y)

$q_2(x,y)$:- R(x,u),R(v,u),R(v,w),R(t,w),R(t,y)

Example of Query Homeomorphism

$var(q_1) \cup const(q_1) = \{x,u, \text{“Smith”}\}$

$var(q_2) = \{x,u,v,w\}$

$q_1(x)$:- R(x,u), R(u,”Smith”), R(u,”Fred”), R(u, u)

$q_2(x)$:- R(x,u), R(u,v), R(u,”Smith”), R(w,u)

The Homeomorphism Theorem

- ***Theorem*** $q_1 \subseteq q_2$ iff there exists a homeomorphism from q_2 to q_1.
- ***Theorem*** Conjunctive query containment is:
  (1) decidable (why ?)
  (2) in NP (why ?)
  (3) NP-hard
- Short: it is NP-complete

Query Equivalence

Definition Query q_1 is equivalent to q_2 if for every database $\mathbf{D}$, $q_1(\mathbf{D}) = q_2(\mathbf{D})$

- Obviously, they are equivalent iff: $q_1 \subseteq q_2$ and $q_2 \subseteq q_1$, hence NP-complete too

Application: Query Minimization

Definition A conjunctive query q is minimal if for any other query q’ equivalent to q, q’ has more goals than q

Query minimization problem: Given q, find an equivalent q_0 that is minimal.

Example of Query Minimization

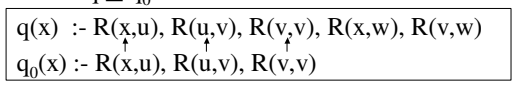
q(x) :- R(x,u), R(u,v), R(v,v), R(x,w), R(v,w)

- Minimal query equivalent to q:

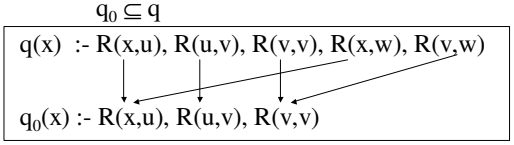
q_0(x) :- R(x,u), R(u,v), R(v,v)

Example of Query Minimization

Proof: $q \subseteq q_0$



$q_0 \subseteq q$



Query Minimization Method

- Given q, find a subset of its goals, q_0 s.t. there exists a homeomorphism from q to q_0
- More practical: for each subgoal, try to see if it can be eliminated.
- Note: all minimal queries equivalent to q are isomorphic (assignment: prove that)

Query Minimization in Practice

q(name) :- Employee(name,age), Employee(name2,age)
Minimizes to:
q_0(name) :- Employee(name,age)

Try this in SQL server:

```
select distinct e1.name
from Employee e1, Employee e2
where e1.age = e2.age
```

Query Minimization in Practice

- Database queries used to be written by programmers
  - "programmers don't write stupid queries"
- Today: queries are often generated automatically
  - E.g. in SilkRoute, XML-QL queries are translated to SQL: NEED MINIMIZATION !