CSE 544: Lecture 10
Theory

Wednesday, April 28, 2004

1

Conjunctive Queries

- A subset of FO queries
- Correspond to SELECT-DISTINCT-FROM-WHERE
- Most queries in practice are conjunctive
- Some optimizers handle only conjunctive queries - break larger queries into many CQs
- CQ's have more positive theoretical properties than arbitrary queries

2

Conjunctive Queries

- **Definition** A conjunctive query is defined by:

$$\varphi ::= R(t_1, \ldots, t_{ar(R)}) \quad | \quad t_i = t_j \quad | \quad \varphi \wedge \varphi' \quad | \quad \exists x.\varphi$$

- missing are $\forall$, $\vee$, $\neg$

- CQ $\subseteq$ FO

3

Conjunctive Queries, CQ

- Example of CQ

$$q(x,y) = \exists z.(R(x,z) \wedge \exists u.(R(z,u) \wedge R(u,y)))$$

$$q(x) = \exists z.\exists u.(R(x,z) \wedge R(z,u) \wedge R(u,y))$$

- Examples of non-CQ:

$$q(x,y) = \forall z.(R(x,z) \text{ à } R(y,z))$$

$$q(x) = T(x) \vee \exists z.S(x,z)$$

4

Conjunctive Queries

- Any CQ query can be written as:

$$q(x_1,\ldots,x_n) = \exists y_1. \exists y_2 \ldots \exists y_p.(R_1(t_{11},\ldots,t_{1m}) \wedge \ldots \wedge R_k(t_{k1},\ldots,t_{km}))$$

(i.e. all quantifiers are at the beginning)

- Same in **Datalog** notation:

Datalog rule

$$q(x_1,\ldots,x_n) \ :- \ R_1(t_{11},\ldots,t_{1m}), \ldots, R_k(t_{k1},\ldots,t_{km}))$$

head — body

5

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)

6

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)

CQs are useful in practice

7

CQ and SQL

CQ:

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

SQL:

Notice "distinct"

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

8

CQ and SQL

- Are CQ queries precisely the SELECT-DISTINCT-FROM-WHERE queries ?

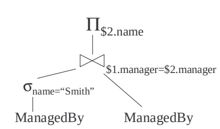
9

CQ and RA

Relational Algebra:

- CQ correspond precisely to σ_C, Π_A, $\times$ (missing: $\cup$, $-$)

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



10

Extensions of CQ

$CQ^{\neq}$

Find managers that manage at least 2 employees

A(y) :- ManagedBy(x,y), ManagedBy(z,y), $x \neq y$

11

Extensions of CQ

$CQ^{<}$

Find employees earning more than their manager:

A(y) :- ManagedBy(x,y), Salary(x,u), Salary(y,v), u>v)

12

Extensions of CQ

CQ¬

Find people sharing the same office with Alice, but not the same manager:

```
A(y) :-  Office("Alice",u), Office(y,u),
         ManagedBy("Alice",x), ¬ManagedBy(x,y)
```

13

Extensions of CQ

UCQ Union of conjuctive queries

Datalog:

```
A(name) :-  Employee(name, dept, age, salary), age > 50
A(name) :-  RetiredEmployee(name, address)
```

Datalog notation is very convenient at expressing unions (no need for $\vee$)

14

Extensions of CQ

- If we extend too much, we capture FO
- Theoreticians need to be careful: small extensions may make a huge difference on certain theoretical properties of CQ

15

Query Equivalence and Containment

- Justified by optimization needs
- Intensively studied since 1977

16

Query Equivalence

- Queries q_1 and q_2 are **equivalent** if for every database **D**, $q_1(\mathbf{D}) = q_2(\mathbf{D})$.
- Notation: $q_1 \equiv q_2$

17

Query Equivalence

```
SELECT x.name, x.manager
FROM    Employee x, Employee y
WHERE x.dept = 'Sales' and x.office = y.office
      and  x.floor = 5 and y.dept = 'Sales'
```

Hmmmm…. Is there a simple way to write that ?

18

Query Containment

- Query 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: $q_1 \subseteq q_2$ and $q_2 \subseteq q_1$ iff $q_1 \equiv q_2$
- Conversely: $q_1 \wedge q_2 \equiv q_2$ iff $q_1 \subseteq q_2$

We will study the containment problem only.

Examples of Query Containments

Is $q_1 \subseteq q_2$?

q_1(x) :- R(x,u), R(u,v), R(v,w)
q_2(x) :- R(x,u), R(u,v)

Examples of Query Containments

Is $q_1 \subseteq q_2$?

q_1(x) :- R(x,u), R(u,v), R(v,x)
q_2(x) :- R(x,u), R(u,x)

Examples of Query Containments

Is $q_1 \subseteq q_2$?

q_1(x) :- R(x,u), R(u,u)
q_2(x) :- R(x,u), R(u,v), R(v,w)

Examples of Query Containments

Is $q_1 \subseteq q_2$?

q_1(x) :- R(x,u), R(u,"Smith")
q_2(x) :- R(x,u), R(u,v)

Query Containment

- **Theorem** Query containment for FO is undecidable
- **Theorem** Query containment for CQ is decidable and NP-complete.

Query Containment Algorithm

How to check $q_1 \subseteq q_2$

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

25

Examples of Canonical Databases

q1(x,y) :- R(x,u),R(v,u),R(v,y)

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

x	u
v	u
v	y

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

26

Examples of Canonical Databases

q1(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

- $t_{q1} = (x)$

27

Checking Containment

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

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	u
v	u
v	y

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

- Yes, $q_1 \subseteq q_2$

28

Query Homomorphisms

- A ***homomorphism*** f : q_2 à q_1 is a function f: var(q_2) à var(q_1) $\cup$ const(q_1) such that:
  - f(body(q_2)) $\subseteq$ body(q_1)
  - $f(t_{q1}) = t_{q2}$

The Homomorphism Theorem $q_1 \subseteq q_2$ iff there exists a homomorphism f : q_2 à q_1

29

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)

Therefore $q_1 \subseteq q_2$

30

Example of Query Homeomorphism

var(q_1) ∪ const(q_1) = {x,u, "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)

Therefore $q_1 \subseteq q_2$

31

The Homeomorphism Theorem

- **Theorem** Conjunctive query containment is:
  - (1) decidable (why ?)
  - (2) in NP (why ?)
  - (3) NP-hard
- Short: it is NP-complete

32

Query Containment for UCQ

$q_1 \cup q_2 \cup q_3 \cup \ldots \subseteq q_1' \cup q_2' \cup q_3' \cup \ldots$

Notice: $q_1 \cup q_2 \cup q_3 \cup \ldots \subseteq q$ iff
$q_1 \subseteq q$ and $q_2 \subseteq q$ and $q_3 \subseteq q$ and ….

Theorem $q \subseteq q_1' \cup q_2' \cup q_3' \cup \ldots$ Iff there exists some k such that $q \subseteq q_k'$

It follows that containment for UCQ is decidable, NP-complete.

33

Query Containment for CQ$^<$

q_1() :- R(x,y), R(y,x)
q_2() :- R(x,y), x < y

$q_1 \subseteq q_2$ although there is no homomorphism !

To check containment do this:
-Consider all possible orderings of variables in q1
-For each of them check containment of q1 in q2
-If all hold, then $q_1 \subseteq q_2$

Still decidable, but harder than NP: now in Π^p_2

34