Exact Minimization of # of Joins

Possible for large set of queries:

- Basic, unnested SQL queries
- One-stage QBE queries
- Relational algebra with π, σ, $\bowtie$
- Relational calculus with $\exists$, $=$, $\wedge$

- Basic idea:

query → pattern (like QBE) → minimize → pattern → simplified query

81

Minimization for a simple case

- Relational algebra with:
  - one input relation
  - π, $\bowtie$
  - σ_{cond} where cond is A = const

↓ rspj query

- First step: write query in QBE style, using a pattern → tableau
  - e.g. tableau for $\pi_{AB}(R) \bowtie \pi_{BC}(R)$, where R: ABC, is:

A	B	C	
a	b	c_1	pattern
a_1	b	c	
a	b	c	answer

82

Tableaux

- Tableau over R: <s, t> where
  - s is the set of rows over R, with variables or constants
  - t is the "answer" row

  t is over a subset of att(R)

  t(A) is a variable occurring in the A column of S, or a constant

- Terminology
  - variables in t: distinguished
    - denoted a, b, c
  - other variables: nondistinguished
    - denoted a_1 a_2 … b_1 b_2 … c_1 c_2

83

Example

- Directors who are also actors

- QBE:

movie	title	dir	actor
		-d	
			-d

answer	dir
I.	-d

- Tableau:

	title	dir	actor	
S	t_1	d	a_1	
	t_2	d_1	d	
T		d		answer row

84

Example

R: ABC

q: $\pi_{AC}(\pi_{AB}(R) \bowtie \pi_{BC}(\sigma_{A=5}(\pi_{AB}(R)) \bowtie \pi_{AC}(R)))$

Relational calculus:

$\exists b_2[\exists c_1(R(ab_2c_1)) \wedge \exists a_1(\exists c_2(R(a_1b_2c_2)) \wedge a_1=5 \wedge \exists b_1(R(a_1b_1c)))]$

Prenex form (move ∃ to left):

$\exists a_1b_1b_2c_1c_2[R(ab_2c_1) \wedge R(a_1b_2c_2) \wedge R(a_1b_1c) \wedge a_1=5]$

85

Example, continued

Tableau:

A	B	C
a	b_2	c_1
5	b_2	c_2
5	b_1	c
a		c

Note: # rows in tableau = 1 + # joins

Note: like QBE query

R	A	B	C
	-a	$-b_2$	
	5	$-b_2$	
	5		-c

answer	A	C
I.	-a	-c

86

Minimizing Tableau

- <s, t>: tableau
- Definition
  - Mapping f on variables is a homomorphism on <s, t> iff:
    - $h(t) = t$
    - $h(c) = c$ if c is constant
    - $h(s) \subseteq s$
- Fact
  - A row $r \in S$ is redundant iff for some homomorphism
    - f of <s, t>, $r \notin f(s)$
  - Note: <f(s), t> equivalent to <s, t>

87

Example

- Tableau <s, t>

A	B	C	
a	b_1	c_1	
a_1	b	c_1	
a	b_2	c_2	s
a_2	b_2	c	
a_2	b_1	c	
a	b	c	t

- Let f: $c_2 \to c_1$
  $b_2 \to b_1$
  $a_1 \to a_1$
  $a_2 \to a_2$

88

Example, continued

- f(<s, t>):

A	B	C
a	b_1	c_1
a_2	b_1	c
a_1	b	c_1
a	b	c

No redundant rows: MINIMAL

- Fact: all minimal equivalent tableaux are isomorphic!

89

Example

R: ABC

q: $\pi_{AB}(\sigma_{B=5}(R)) \bowtie \pi_{BC}(\pi_{AB}(R) \bowtie \pi_{AC}(\sigma_{B=5}(R)))$

Tableau:

A	B	C
a	5	c_1
a_1	5	c_2
a_1	5	c
a	5	c

Minimization
$c_2 \rightarrow c$

↓

A	B	C
a	5	c_1
a_1	5	c
a	5	c

Minimized query

↓

$\sigma_{B=5}(\pi_{AB}(R) \bowtie \pi_{BC}(R))$

90

Functional Dependencies

- Dependencies: statements about properties of valid data
  - e.g.: “Every student is a person”
    - inclusion dependency
  - “Each employee works in more than one department”
    - NAME → DEPARTMENT
    - functional dependency
- Use of dependencies:
  - check data integrity
  - query optimization
  - schema design → “normal forms”

Functional Dependencies

- Functional dependency over R:
  - expression $x \rightarrow y$ where $x, y \subseteq att(R)$
- A relation r over R satisfies $x \rightarrow y$
  - iff whenever two tuples in y agree on x, they also agree on y

e.g.

SCHEDULE	THEATER	TITLE
	odeon	tango
	forum	delicatessen

Satisfies THEATER → TITLE

SCHEDULE	THEATER	TITLE
	odeon	tango
	forum	delicatessen
	forum	l’amant

Violates THEATER → TITLE, satisfies TITLE → THEATER

Using FDs in Query Optimization

- Example: R: ABC with $B \rightarrow C$
  - query $\pi_{AB}(R) \bowtie \pi_{BC}(R)$
- Fact: if R satisfies $B \rightarrow C$ then
  - $\pi_{AB}(R) \bowtie \pi_{BC}(R) = R$
  - Why: tableau of query is

A	B	C
a	b	c_1
a_1	b	c
a	b	c

  - Note: if <a, b, c> ∈ R and <a_1, b, c> ∈ R then $c_1 = c$ since R satisfies $B \rightarrow C$
  - So, tableau is equivalent to

A	B	C
a	b	c
a	b	c

≡ R

- In general: can simplify tableau <s, t> over R if R satisfies a set F of FDs.
- Algorithm: The Chase
  - Input: tableau <s, t>, set F of FDs
  - Output: tableau $CHASE_F$<s, t> on all relations satisfying F
- Note: assume without loss of generality that FDs in F are of the form $X \rightarrow A$ where A is one attribute

The Chase

- Repeat until no change
  - For each $X \rightarrow A$ in F do
    - For each t_1, t_2 in s such that $t_1(x) = t_2(x)$, $t_1(A) \neq t_2(A)$ do
      - if $t_1(A)$, $t_2(A)$ are nondistinguished then replace one by the other in s
      - if $t_1(A)$ distinguished, $t_2(A)$ nondistinguished then replace $t_2(A)$ by $t_1(A)$ in s
      - if $t_1(A)$ is constant, $t_2(A)$ is variable then replace $t_2(A)$ by $t_1(A)$ in s
      - if $t_1(A)$ is constant, $t_2(A)$ is constant then STOP and output ∅

95

Optimization of RSPJ Queries with FDs

- q over R, set of FDs F over R
  - build tableau <s, t> of q
  - compute $CHASE_F$<s, t>
  - minimize $CHASE_F$<s, t>
  - construct rspj query from minimal tableau

96

Example

– R: ABC   F = {B → A}
– $q = \pi_{BC}(\sigma_{A=5}(R)) \rhd\lhd \pi_{AB}(R)$

<s, t>:

A	B	C
5	b	c
a	b	c_1
a	b	c

CHASE<s, t>:

A	B	C
5	b	c
5	b	c_1
5	b	c

MIN:

A	B	C
5	b	c
5	b	c

RSPJ: $\sigma_{A=5}(R)$

97

Example

– R: ABC   F = {B → A}, $q = \pi_{BC}(\sigma_{A=5}(R)) \rhd\lhd \pi_{AB}(\sigma_{A=6}(R))$

<s, t>:

A	B	C
5	b	c
6	b	c_1
6	b	c

CHASE<s, t>: ∅

QUERY: ∅

98

Example

– R: ABC F = {A → B}, $q = \pi_{AB}(R) \bowtie \pi_A(\sigma_{B=5}(R)) \bowtie \pi_{AB}(\pi_{AC}(R) \bowtie \pi_{BC}(R))$

<s, t>:

A	B	C
a	b	c_1
a	b_1	c_2
a_1	b	c_2
a	5	c_3
a	b	

CHASE<s, t>:

A	B	C
a	5	c_1
a	5	c_2
a_1	5	c_2
a	5	c_3
a	5	

MIN:

A	B	C
a	5	c_1
a	5	

RSPJ: $\pi_{AB}(\sigma_{B=5}(R))$

99