CS 450

Relational Algebra 2

Prof. Jessica Lin

Relational Algebra (So far)

- Basic operations:
  - *Selection* (σ) Selects a subset of rows from relation.
  - *Projection* (π) Deletes unwanted columns from relation.
  - *Cross-product* ($\times$) Allows us to combine two relations.
  - *Set-difference* (-) Tuples in reln. 1, but not in reln. 2.
  - *Union* ($\cup$) Tuples in reln. 1 and tuples in reln. 2.

  Also,
  - *Rename* (ρ) Changes names of the attributes
  - *Intersection* ($\cap$) Tuples in both reln. 1 and in reln. 2.

- Since each operation returns a relation, operations can be *composed*! (Algebra is "closed".)
- Use of temporary relations recommended.

Another example

- Find the name of the sailor having the highest rating.

$$\text{AllR} \leftarrow \pi_{rating} S2$$

$$\text{Result?} = \pi_{Sname}(\sigma_{S2.rating<AllR.rating}(S2 \times \text{AllR}))$$

What's in "Result?" ?

Does it answer our query?

S2

sid	sname	rating	age
28	yuppy	9	35.0
31	lubber	8	55.5
44	guppy	5	35.0
58	rusty	10	35.0

S2

sid	sname	rating	age
28	yuppy	9	35.0
31	lubber	8	55.5
44	guppy	5	35.0
58	rusty	10	35.0

×

AllR

rating
9
8
5
10

=

sid	sname	S2.rating	age	AllR.rating
28	yuppy	9	35.0	9
28	yuppy	9	35.0	8
28	yuppy	9	35.0	5
28	**yuppy**	**9**	**35.0**	**10**
31	**lubber**	**8**	**55.5**	**9**
31	lubber	8	55.5	8
31	lubber	8	55.5	5
31	**lubber**	**8**	**55.5**	**10**
44	**guppy**	**5**	**35.0**	**9**
44	**guppy**	**5**	**35.0**	**8**
44	guppy	5	35.0	5
44	**guppy**	**5**	**35.0**	**10**
58	rusty	10	35.0	9
58	rusty	10	35.0	8
58	rusty	10	35.0	5
58	rusty	10	35.0	10

$\text{AllR} \leftarrow \pi_{rating} S2$

$\text{Result?} = \pi_{Sname}(\sigma_{S2.rating < AllR.rating}(S2 \times \text{AllR}))$

This doesn’t work. We need to consider something else.

Review: Union, Intersection, Set-Difference

- All of these operations take two input relations, which must be *compatible*:
  - Same number of fields.
  - 'Corresponding' fields have the same type.
- What is the *schema* of result?

sid	sname	rating	age
22	dustin	7	45.0
31	lubber	8	55.5
58	rusty	10	35.0
44	guppy	5	35.0
28	yuppy	9	35.0

$S1 \cup S2$

sid	sname	rating	age
22	dustin	7	45.0

$S1 - S2$

sid	sname	rating	age
31	lubber	8	55.5
58	rusty	10	35.0

$S1 \cap S2$

Back to our query

- Find the name of the sailor having the highest rating.

$$\text{AllR} \leftarrow \pi_{rating} S2$$

$$Tmp \leftarrow \pi_{Sid,Sname}(\sigma_{S2.rating<AllR.rating}(S2 \times \text{AllR}))$$

$$\text{Result} = \pi_{Sname}(\pi_{Sid,Sname}(S2) - \text{Tmp})$$

* Why not project on Sid only for Tmp?

Additional Operations

We define additional operations that do not add any power to the relational algebra, but that simplify common queries.

- Natural join
- Conditional Join
- Equi-Join
- Division

All joins are really special cases of conditional join

Also, we've already seen "Set intersection":

$$r \cap s = r - (r - s)$$

Quick note on notation

good_customers

customer-name	***loan-number***
Patty	1234
Apu	3421
Selma	2342
Ned	4531

bad_customers

customer-name	***loan-number***
Seymour	3432
Marge	3467
Selma	7625
Abraham	3597

If we have two or more relations which feature the same attribute names, we could confuse them. To prevent this we can use dot notation.
For example

good_customers.loan-number

Natural-Join Operation: Motivation

Very often, we have a query and the answer is not contained in a single relation. For example, I might wish to know where Apu banks.
The classic relational algebra way to do such queries is a cross product, followed by a selection which tests for equality on some pair of fields.

$\sigma_{borrower.l\text{-}number\ =\ loan.l\text{-}number}(borrower \times loan)))$

While this works…

- it is unintuitive
- it requires a lot of memory
- the notation is cumbersome

borrower

cust-name	***l-number***
Patty	1234
Apu	3421

loan

l-number	***branch***
1234	Dublin
3421	Irvine

cust-name	***borrower.l-number***	***loan.l-number***	***branch***
Patty	1234	1234	Dublin
Patty	1234	3421	Irvine
Apu	3421	1234	Dublin
Apu	3421	3421	Irvine

cust-name	***borrower.l-number***	***loan.l-number***	***branch***
Patty	1234	1234	Dublin
Apu	3421	3421	Irvine

Note that in this example the two relations are the same size (2 by 2), this does not have to be the case.

So, we have a more intuitive way of achieving the same effect, the natural join, denoted by the $\bowtie$ symbol

Natural-Join Operation: Intuition

Natural join combines a cross product and a selection into one operation. It performs a selection forcing equality on *those attributes that appear in both relation schemes*. Duplicates are removed as in all relation operations.

So, if the relations have one attribute in common, as in the last slide ("*l-number*"), for example, we have…

$$borrower \bowtie loan \quad = \sigma_{borrower.l\text{-}number\, =\, loan.l\text{-}number}(borrower \times loan)))$$

There are two special cases:

- If the two relations have no attributes in common, then their natural join is simply their cross product.
- If the two relations have more than one attribute in common, then the natural join selects only the rows where all pairs of matching attributes match. (let's see an example on the next slide).

A

l-name	*f-name*	*age*
Bouvier	Selma	40
Bouvier	Patty	40
Smith	Maggie	2

B

l-name	*f-name*	*ID*
Bouvier	Selma	1232
Smith	Selma	4423

Both the *l-name* and the *f-name* match, so select.

Only the *f-names* match, so don't select.

Only the *l-names* match, so don't select.

l-name	*f-name*	*age*	*l-name*	*f-name*	*ID*
Bouvier	**Selma**	40	**Bouvier**	**Selma**	1232
Bouvier	**Selma**	40	Smith	**Selma**	4423
Bouvier	Patty	2	**Bouvier**	Selma	1232
Bouvier	Patty	40	Smith	Selma	4423
Smith	Maggie	2	Bouvier	Selma	1232
Smith	Maggie	2	**Smith**	Selma	4423

We remove duplicate attributes…

l-name	*f-name*	*age*	*l-name*	*f-name*	*ID*
Bouvier	Selma	40	Bouvier	Selma	1232

The natural join of *A* and *B*

Note that this is just a way to visualize the natural join, we don't really have to do the cross product as in this example

$A \bowtie B =$

l-name	*f-name*	*age*	*ID*
Bouvier	Selma	40	1232

Natural-Join Operation

- Notation: $r \bowtie s$
- Let r and s be relation instances on schemas R and S respectively.The result is a relation on schema $R \cup S$ which is obtained by considering each pair of tuples t_r from r and t_s from s.
- If t_r and t_s have the same value on each of the attributes in $R \cap S$, a tuple t is added to the result, where
  - t has the same value as t_r on r
  - t has the same value as t_s on s
- Example:

  $R = (A, B, C, D)$

  $S = (E, B, D)$
- Result schema = (A, B, C, D, E)
- $r \bowtie s$ is defined as:

$$\pi_{r.A,\ r.B,\ r.C,\ r.D,\ s.E}\ (\sigma_{r.B\ =\ s.B\ \ r.D\ =\ s.D}\ (r\ \mathrm{x}\ \ s))$$

Natural Join Operation – Example

- Relation instances *r, s*:

A	B	C	D
α	1	α	a
β	2	γ	a
γ	4	β	b
α	1	γ	a
δ	2	β	b

r

B	D	E
1	a	α
3	a	β
1	a	γ
2	b	δ
3	b	$\in$

s

$r \bowtie s$

A	B	C	D	E
α	1	α	a	α
α	1	α	a	γ
α	1	γ	a	α
α	1	γ	a	γ
δ	2	β	b	δ

How did we get here?

Lets do a trace over the next few slides…

A	B	C	D
α	1	α	a
β	2	γ	a
γ	4	β	b
α	1	γ	a
δ	2	β	b

r

B	D	E
1	a	α
3	a	β
1	a	γ
2	b	δ
3	b	∈

s

First we note which attributes the two relations have in common…

A	B	C	D
α	1	α	a
β	2	γ	a
γ	4	β	b
α	1	γ	a
δ	2	β	b

r

B	D	E
1	a	α
3	a	β
1	a	γ
2	b	δ
3	b	∈

s

A	B	C	D	E
α	1	α	a	α
α	1	α	a	γ

There are two rows in s that match our first row in r, (in the relevant attributes) so both are joined to our first row…

A	B	C	D
α	1	α	a
β	2	γ	a
γ	4	β	b
α	1	γ	a
δ	2	β	b

r

B	D	E
1	a	α
3	a	β
1	a	γ
2	b	δ
3	b	∈

s

A	B	C	D	E
α	1	α	a	α
α	1	α	a	γ

…there are no rows in s that match our second row in r, so do nothing…

A	B	C	D
α	1	α	a
β	2	γ	a
γ	4	β	b
α	1	γ	a
δ	2	β	b

r

B	D	E
1	a	α
3	a	β
1	a	γ
2	b	δ
3	b	∈

s

A	B	C	D	E
α	1	α	a	α
α	1	α	a	γ

…there are no rows in s that match our third row in r, so do nothing…

A	B	C	D
α	1	α	a
β	2	γ	a
γ	4	β	b
α	1	γ	a
δ	2	β	b

r

B	D	E
1	a	α
3	a	β
1	a	γ
2	b	δ
3	b	∈

s

A	B	C	D	E
α	1	α	a	α
α	1	α	a	γ
α	1	γ	a	α
α	1	γ	a	γ

There are two rows in s that match our fourth row in r, so both are joined to our fourth row…

A	B	C	D
α	1	α	a
β	2	γ	a
γ	4	β	b
α	1	γ	a
δ	2	β	b

r

B	D	E
1	a	α
3	a	β
1	a	γ
2	b	δ
3	b	∈

s

A	B	C	D	E
α	1	α	a	α
α	1	α	a	γ
α	1	γ	a	α
α	1	γ	a	γ
δ	2	β	b	δ

There is one row that matches our fifth row in r,.. so it is joined to our fifth row and we are done!

Natural Join on Sailors Example

S1

<u>sid</u>	sname	rating	age
22	dustin	7	45.0
31	lubber	8	55.5
58	rusty	10	35.0

R1

<u>sid</u>	<u>bid</u>	<u>day</u>
22	101	10/10/96
58	103	11/12/96

$S1 \bowtie R1 =$

sid	sname	rating	age	bid	day
22	dustin	7	45.0	101	10/10/96
58	rusty	10	35.0	103	11/12/96

Earlier We Saw…

Query: Find the name of the sailor who reserved boat 101.

$$Temp \leftarrow \rho_{S1\_New(sid \rightarrow sid1)}(S1) \times \rho_{R1\_New(sid->sid2)}(R1)$$

$$\text{Result} = \pi_{Sname}(\sigma_{sid1=sid2 \ \wedge \ bid=101}(Temp))$$

* Note my use of "temporary" relation Temp.

Query revisited using natural join

Query: Find the name of the sailor who reserved boat 101.

$$\text{Result} = \pi_{Sname}(\sigma_{bid=101}(S1 \bowtie R1))$$

Or

$$\text{Result} = \pi_{Sname}(S1 \bowtie \sigma_{bid=101}(R1))$$

What's the difference between these two approaches?

Conditional-Join Operation:

The conditional join is actually the most general type of join. I introduced the natural join first only because it is more intuitive and... natural!

Just like natural join, conditional join combines a cross product and a selection into one operation. However instead of only selecting rows that have equality on those attributes that appear in both relation schemes, we allow selection based on any predicate.

$$r \bowtie_c s = \sigma_c(r \times s)$$

Where c is any predicate the attributes of r and/or s

Duplicate rows are removed as always, but duplicate columns are not removed!

Conditional-Join Example:

We want to find all women that are younger than their husbands…

r

l-name	*f-name*	*marr-Lic*	*age*
Simpson	Marge	777	35
Lovejoy	Helen	234	38
Flanders	Maude	555	24
Krabappel	Edna	978	40

s

l-name	*f-name*	*marr-Lic*	*age*
Simpson	Homer	777	36
Lovejoy	Timothy	234	36
Simpson	Bart	*null*	9

$$r \bowtie_{r.age < s.age \text{ AND } r.Marr\text{-}Lic = s.Marr\text{-}Lic} s$$

r.l-name	*r.f-name*	*r.Marr-Lic*	*r.age*	*s.l-name*	*s.f-name*	*s.marr-Lic*	*s.age*
Simpson	Marge	777	35	Simpson	Homer	777	36

Note we have removed ambiguity of attribute names by using "dot" notation
Also note the redundant information in the *marr-lic* attributes

Equi-Join

- Equi-Join: Special case of conditional join where the conditions consist only of equalities.
- Natural Join: Special case of equi-join in which equalities are specified on ALL fields having the same names in both relations.

Equi-Join

r

l-name	*f-name*	*marr-Lic*	*age*
Simpson	Marge	777	35
Lovejoy	Helen	234	38
Flanders	Maude	555	24
Krabappel	Edna	978	40

s

l-name	*f-name*	*marr-Lic*	*age*
Simpson	Homer	777	36
Lovejoy	Timothy	234	36
Simpson	Bart	*null*	9

$$r \bowtie_{r.Marr\text{-}Lic \,=\, s.Marr\text{-}Lic} s$$

r.l-name	*r.f-name*	*Marr-Lic*	*r.age*	*s.l-name*	*s.f-name*	*s.age*
Simpson	Marge	777	35	Simpson	Homer	36
Lovejoy	Helen	234	38	Lovejoy	Timothy	36

Review on Joins

- All joins combine a cross product and a selection into one operation.
- Conditional Join
  - the selection condition can be of any predicate (e.g. rating1 > rating2)
- Equi-Join:
  - Special case of conditional join where the conditions consist only of equalities.
- Natural Join
  - Special case of equi-join in which equalities are specified on ALL fields having the same names in both relations.

A Note on Precedence

- Unary operators have the highest precedence: [σ, π, ρ]
- Then “multiplicative” operators: [×, ⋈]
- Then “additive” operators: [∩, ∪, −]

Banking Examples

branch (branch-id, branch-city, assets)

customer (customer-id, customer-name, customer-city)

account (account-number, branch-id, balance)

loan (loan-number, branch-id, amount)

depositor (customer-id, account-number)

borrower (customer-id, loan-number)

Example Queries 1

- Find all loans over $1200

"select from the relation *loan*, only the rows which have a *amount* greater than 1200"

loan

loan-number	***branch-id***	***amount***
1234	001	1,923.03
3421	002	123.00
2342	004	56.25
4531	005	120.03

$\sigma_{amount > 1200}(loan)$

1234	001	1,923.03

Example Queries 2

- Find the loan number for each loan of an amount greater than $1200

"select from the relation *loan*, only the rows which have a *amount* greater than 1200, then project out just the *loan_number*"

loan

loan-number	*branch-id*	*amount*
1234	001	1,923.03
3421	002	123.00
2342	004	56.25
4531	005	120.03

$\sigma_{amount > 1200}(loan)$

1234	001	1,923.03

$\pi_{loan\text{-}number}(\sigma_{amount > 1200}(loan))$

1234

Example Queries 3

- Find all loans greater than \$1200 or less than \$75

"select from the relation *loan*, only the rows which have a *amount* greater than 1200 or an *amount* less than 75

loan

loan-number	*branch-id*	*amount*
1234	001	1,923.03
3421	002	123.00
2342	004	56.25
4531	005	120.03

$\sigma_{amount > 1200 \vee amount < 75}(loan)$

1234	001	1,923.03
2342	004	56.25

Example Queries 4

- Find the IDs of all customers who have a loan, an account, or both, from the bank

borrower

customer-id	***loan-number***
201	1234
304	3421
425	2342
109	4531

depositor

customer-id	***account-number***
333	3467
304	2312
201	9999
492	3423

$\pi_{customer\text{-}id}$ (*borrower*)

201
304
425
109

$\pi_{customer\text{-}id}$ (*depositor*)

333
304
201
492

201
304
425
109
333
492

$\pi_{customer\text{-}id}$ (*borrower*) $\cup$ $\pi_{customer\text{-}id}$ (*depositor*)

Example Queries 5

Note this example is split over two slides!

Find the IDs of all customers who have a loan at branch 001.

We retrieve *borrower* and *loan*…

borrower

customer-id	***loan-number***
201	1234
304	3421

loan

loan-number	***branch-id***	***amount***
1234	001	1,923.03
3421	002	123.00

…we calculate their cross product…

customer-id	***borrower.loan-number***	***loan.loan-number***	***branch-id***	***amount***
201	1234	1234	001	1,923.03
201	1234	3421	002	123.00
304	3421	1234	001	1,923.03
304	3421	3421	002	123.00

…we calculate their cross product…

customer-id	*borrower.loan-number*	*loan.loan-number*	*branch-id*	*amount*
201	1234	1234	001	1,923.03
201	1234	3421	002	123.00
304	3421	1234	001	1,923.03
304	3421	3421	002	123.00

…we select the rows where *borrower.loan-number* is equal to *loan.loan-number*…

customer-id	*borrower.loan-number*	*loan.loan-number*	*branch-id*	*amount*
201	1234	1234	001	1,923.03
304	3421	3421	002	123.00

…we select the rows where *branch-id* is equal to *"001"*

customer-id	*borrower.loan-number*	*loan.loan-number*	*branch-id*	*amount*
201	1234	1234	001	1,923.03

…we project out the *customer-id.*

201

$$\pi_{customer\text{-}id}\,(\sigma_{branch\text{-}id=\text{'}001\text{'}}\,(\sigma_{borrower.loan\text{-}number\,=\,loan.loan\text{-}number}(borrower \times loan)))$$

Now Using Natural Join

Find the IDs of all customers who have a loan at branch 001.

We retrieve *borrower* and *loan*…

borrower

customer-id	***loan-number***
201	1234
304	3421

loan

loan-number	***branch-id***	***amount***
1234	001	1,923.03
3421	002	123.00

1234 in *borrower* is matched with 1234 in *loan*…

3421 in *borrower* is matched with 3421 in *loan*…

customer-id	***loan-number***	***branch-id***	***amount***
201	1234	001	1,923.03
304	3421	002	123.00

The rest is the same.

customer-id	***loan-number***	***branch-id***	***amount***
201	1234	001	1,923.03

$$\pi_{customer\text{-}id}\,(\sigma_{branch\text{-}id='001'}\,(\sigma_{borrower.loan\text{-}number\,=\,loan.loan\text{-}number}(borrower \times loan)))$$

$$= \quad \pi_{customer\text{-}id}\,(\sigma_{branch\text{-}id='001'}\,(borrower \bowtie loan))$$

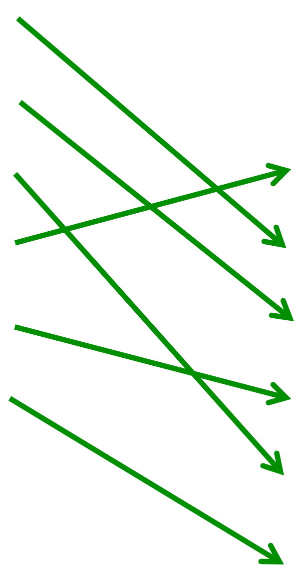
Example Queries 7

Note this example is split over two slides!

- Find the *names* of all customers who have a loan, an account, or both, from the bank

$\pi_{customer\text{-}id}\,(borrower) \cup \pi_{customer\text{-}id}\,(depositor)$

201
304
425
109
333
492



customer

customer-id	***customer-name***	***customer-city***
101	Carol	Fairfax
109	David	Fairfax
201	John	Vienna
304	Mary	McLean
333	Ben	Chantilly
425	David	Manassas
492	Jason	Fairfax
501	Adam	Burke

Example Queries

- Find the *names* of all customers who have a loan, an account, or both, from the bank

customer-id	*customer-name*	*customer-city*
109	David	Fairfax
201	John	Vienna
304	Mary	McLean
333	Ben	Chantilly
425	David	Manassas
492	Jason	Fairfax

customer-name
David
John
Mary
Ben
David
Jason

customer-name
David
John
Mary
Ben
Jason

$$\pi_{customer\text{-}name}((\pi_{customer\text{-}id}(borrower) \cup \pi_{customer\text{-}id}(depositor)) \bowtie \text{customer})$$

Example Queries 8

Note this example is split over three slides!

Find the largest account balance

We do a rename to get a "copy" of the balance column from *account*. We call this copy *d*…

… next we will do a cross product…

account

account-number	***branch-id***	***balance***
7777	001	100.30
8888	003	12.34
6666	004	45.34

d

balance
100.30
12.34
45.34

… do a cross product…

account-number	*branch-id*	*account. balance*	*d.balance*
7777	001	100.30	100.30
7777	001	100.30	12.34
7777	001	100.30	45.34
8888	003	12.34	100.30
8888	003	12.34	12.34
8888	003	12.34	45.34
6666	004	45.34	100.30
6666	004	45.34	12.34
6666	004	45.34	45.34

…select out all rows where *account.balance* is less than *d.balance*…

account-number	*branch-id*	*account. balance*	*d.balance*
8888	003	12.34	100.30
8888	003	12.34	45.34
6666	004	45.34	100.30

.. next we project out *account.balance*…

…then we do a set difference between it and the original *account.balance* from the account relation…

… the set difference leaves us with one number, the largest value!

account-number	***branch-id***	***account.balance***	***d.balance***
8888	003	12.34	100.30
8888	003	12.34	45.34
6666	004	45.34	100.30

balance from account

balance
100.30
12.34
45.34

—

account.balance
12.34
45.34

100.30

$\pi_{balance}(account) - \pi_{account.balance}(\sigma_{account.balance < d.balance}(account \times \rho_d(\pi_{balance}(account))))$

Now Using Conditional Join

Find the largest account balance

$\pi_{balance}(account) - \pi_{account.balance}(\sigma_{account.balance < d.balance}(account \times \rho_d(\pi_{balance}(account))))$

$\rho_d(\pi_{balance}(account))$

$\pi_{balance}(account) - \pi_{account.balance}(account \bowtie_{account.balance < d.balance} d)$