

Kibővített relációs algebra

Tankönyv: Ullman-Widom:
Adatbázisrendszerek Alapvetés
Második, átdolgozott kiadás,
Panem, 2009

- 5.1. Relációs műveletek
multihalmazokon
- 5.2. Kiterjesztett műveletek
a relációs algebrában



Forrás: J.D.Ullman slides, és Benczúr András Adatbázisok-1 előadása

Relációs algebra kibővítése

- Az eddig tanult műveleteket: **vetítés** (Π), **kiválasztás** (σ), **halmazműveletek**: **unió** ($\cup$), **különbség** ($-$), **metszet** ($\cap$), **szorzás**: **természetes összekapcsolás** ($\bowtie$), **direkt-szorzat** ($\times$), stb. **multihalmazok fölött értelmezzük, mint az SQL-ben**, egy reláció nem sorok halmazából, hanem **multihalmazából** áll, vagyis megengedett a sorok ismétlődése.
- Ezeken kívül a SELECT **kiegészítéseinek és záradékainak** megfeleltetett **új műveletekkel is kibővítjük** a rel. algebrát:
 - **Ismétlődések megszüntetése** (δ) - **select distinct ..**
 - **Összesítő műveletek és csoportosítás** (γ_{lista}) - **group by..**
 - **Vetítési művelet kiterjesztése** (Π_{lista}) - **select kif [as onev]..**
 - **Rendezési művelet** (τ_{lista}) - **order by..**
 - **Külső összekapcsolások** ($\overset{O}{\bowtie}$) – **[left | right | full] outer join**

Multihalmazok egyesítése, metszete, különbsége

- **Unió:** $R \cup S$ -ben egy t sor annyiszor fordul elő ahányszor előfordul R -ben, plusz ahányszor előfordul S -ben: $n+m$
- **Metszet:** $R \cap S$ -ben egy t sor annyiszor fordul elő, amennyi az R -ben és S -ben lévő előfordulások minimuma: $\min[n, m]$
- **Különbség:** $R - S$ -ben egy t sor annyiszor fordul elő, mint az R -beli előfordulások mínusz az S -beli előfordulások száma, ha ez pozitív, egyébként pedig 0, vagyis $\max[0, n-m]$

R			S			A		B	
A	B		A	B		A		B	
1	3	$\cup$	1	3	$\Rightarrow$	1		3	
1	2		2	5		1		2	
						1		3	
						2		5	

A „többi művelet” multihalmazok fölött

- A projekció, szelekció, Descartes-szorzat, természetes összekapcsolás és Théta-összekapcsolás végrehajtása során nem küszöböljük ki az ismétlődéseket.

R			$\Pi_A(R)$	
A	B		A	
1	2	→	1	
1	5		1	
2	3		2	

Ismétlődések megszüntetése duplikátumok kiszűrése

- **Ismétlődések megszüntetése:** $R1 := \delta(R2)$
- A művelet jelentése: R2 multihalmazból R1 halmazt állít elő, vagyis az R2-ben egyszer vagy többször előforduló sorok csak egyszer szerepelnek az R1-ben.
- A DISTINCT reprezentálására szolgál

$R = ($

A	B
1	2
3	4
1	2

$\delta(R) =$

A	B
1	2
3	4

Összesítő függvények

R =

A	B
1	3
3	4
3	2

SUM(A) = 7

COUNT(A) = 3

MIN(B) = 2

MAX(B) = 4

AVG(B) = 3

Csoportosítás és összesítés $\gamma_L(R)$

- A művelet a **csoportosítást** (**GROUP BY**), a csoportokon végezhető **összesítő függvényeket** (AVG, SUM, COUNT, MIN, MAX) reprezentálja.
- A művelet jele: $\gamma_L(R)$.
- Itt az **L** lista valamennyi eleme a következők egyike:
 - R egy attribútuma: ez az attribútum egyike a csoportosító attribútumoknak (a GROUP BY után jelenik meg).
 - a reláció egyik attribútumára (ez az összesítési attribútumra) alkalmazott összesítő operátor, ha az összesítés eredményére névvel szeretnénk hivatkozni, akkor nyilat és új nevet használunk.

Csoportosítás és összesítés $\gamma_L(R)$

- Osszuk R sorait csoportokba. Egy csoport azokat a sorokat tartalmazza, amelyek az L listán szereplő csoportosítási attribútumokhoz tartozó értékei megegyeznek
 - Vagyis ezen attribútumok minden egyes különböző értéke egy csoportot alkot.
- Minden egyes csoporthoz számoljuk ki az L lista összesítési attribútumaira vonatkozó összesítéseket
- Az eredmény minden egyes csoportra egy sor:
 1. A csoportosítási attribútumok és
 2. Az összesítési attribútumra vonatkozó összesítések (az adott csoport összes sorára)

Példa: Csoportosításra és összesítésre

R =

A	B	C
1	2	3
4	5	6
1	2	5

$\gamma_{A,B,AVG(C)} \rightarrow X (R) = ??$

Először csoportosítunk

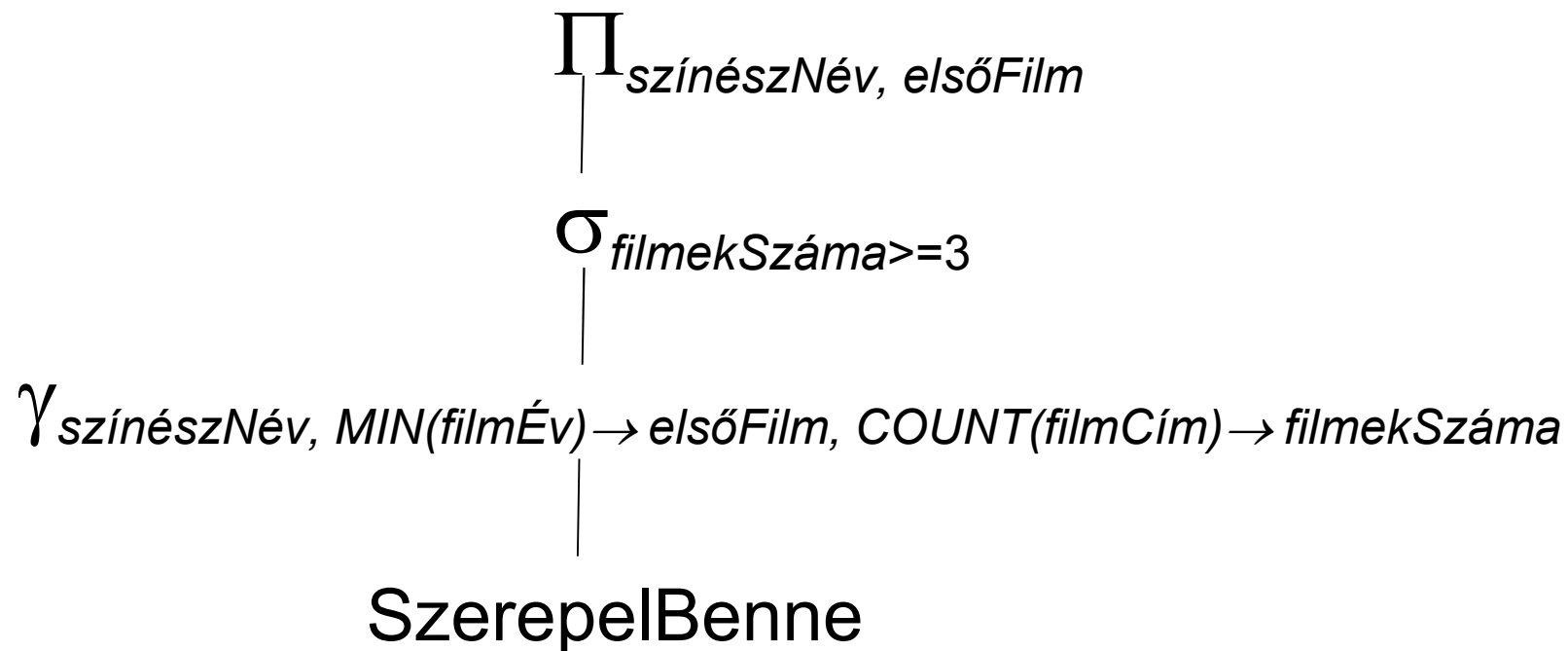
A	B	C
1	2	3
1	2	5
4	5	6

majd csoportonként
összesítünk:

A	B	X
1	2	4
4	5	6

Tankönyv példája

- Adjuk meg azokat a színészeket, akik már szerepeltek legalább három filmben illetve adjuk meg az első olyan filmet is, amiben szerepelt:



Általánosított vetítés

- $\Pi_L(R)$ L listája az SQL **SELECT** záradékához hasonlóan tartalmazhatja:
 - R egy attribútumát,
 - $x \rightarrow y$, ahol x, y attribútumnevek, s itt x-et y-ra nevezzük át,
 - $E \rightarrow y$, ahol E R attribútumait, konstansokat, aritmetikai operátorokat és karakterlánc operátorokat tartalmazhat például: $A + 5$, $C \parallel 'nevű emberek'$.

R			$\Pi_{A,B+C \rightarrow x}(R)$	
A	B	C	A	x
0	1	3	0	4
4	9	1	4	10
5	5	7	5	12



Rendezés

- **Rendezés:** $\tau_{A_1, \dots, A_n}(R)$.
- Először A_1 attribútum szerint rendezzük R sorait. Majd azokat a sorokat, amelyek értéke megegyezik az A_1 attribútumon, A_2 szerint, és így tovább.
- Hasonlóan az **ORDER BY** működéséhez.
- Ez az egyetlen olyan művelet, amelynek az eredménye se nem halmaz se nem multihalmaz.

$R =$

(A	B)
1	2
3	4
5	2

$$\tau_B(R) = [(5,2), (1,2), (3,4)]$$

Külső összekapcsolások

- Ez **nem relációs algebrai művelet**, uis kilép a modellből.
- Lehet baloldali, jobboldali, teljes külső összekapcsolás.
- R, S sémái $R(A_1, \dots, A_n, B_1, \dots, B_k)$, ill. $S(B_1, \dots, B_k, C_1, \dots, C_m)$
- $R \bowtie S = R \bowtie S$ relációt kiegészítjük az R és S soraival, a hiányzó helyekre NULL értéket írva megőrzi a „lógó sorokat”

Példa: külső összekapcsolás

$R =$ (

A	B
1	2
4	5

)

$S =$ (

B	C
2	3
6	7

)

(1,2) és (2,3) összekapcsolható, a másik két sor azonban „lóg”.

$R \text{ OUTERJOIN } S =$

A	B	C
1	2	3
4	5	NULL
NULL	6	7

Folytatás...

- Visszatérünk az SQL 1999-es szabványban a FROM záradékban az összekapcsolások új szintaxisára és a külső összekapcsolásokra.