5.2. Kiterjesztett műveletek a relációs algebrában

A klasszikus relációs algebrát a 2.4. alfejezetben már bemutattuk, illetve az olyan módosításokat is megtettük az 5.1. alfejezetben, amelyek elengedhetetlenek voltak ahhoz, hogy a korábbi halmazos szemlélettel szemben most a relációkra, mint sorok multihalmazára tekinthessünk. Ennek a két résznek az elképzelései képezik a legtöbb korszerű lekérdező nyelv alapjait. Habár az olyan nyelvek, mint az SQL, számos olyan, ezektől eltérő műveletet is tartalmaznak, amelyek alkalmazások írásánál nagyon fontosak lehetnek. Ezért a relációs műveletek egy teljesebb kezelési megközelítésében szerepelni kell még néhány más műveletnek is, amelyeket ezen alrész során be is fogunk vezetni. A kiegészítések a következők:

1. *Az ismétlődések megszüntetésének művelete: δ*. Ez a művelet a multihalmazt halmazzá alakítja a sorok másolatainak megszüntetésével.

2. *Az összesítő műveletek*. Ilyen például az összegzés, illetve az átlag, amelyek ugyan nem a relációs algebra műveletei, de csoportosító műveletekkel használhatók (ezt később részletezzük). Az összesítő műveletek egy reláció attribútumaira (oszlopaira) vannak értelmezve. Például egy oszlopra vonatkozó összegzés értéke azt a számot reprezentálja, amelyet az oszlopban szereplő értékek összeadása által kapunk meg.

3. *Csoportosítás*. A sorok csoportosítása egy reláció sorainak „csoportokba" történő besorolása a reláció egy vagy több attribútumának értékétől függően. Ezek után már az egyes csoportok oszlopaira összesítési művelet is végezhető, amely lehetővé teszi számunkra néhány olyan lekérdezés megfogalmazását is, amelyeket a klasszikus relációs algebrával nem tudtunk kifejezni. A γ *csoportosítási művelet* egy olyan művelet, amely a csoportosítás és az összesítés hatását kombinálja.

4. *Kiterjesztett vetítés művelet.* Ez kiterjeszti a π művelet hatását azzal, hogy néhány oszlopra történő vetítés mellett azt is lehetővé teszi, hogy az érintett oszlopok valamilyen összesítési relációja alapján új oszlopok kiszámítását is elvégezhessük.

5. *Rendezési művelet*. A τ egy relációt a sorainak egy vagy több attribútumtól függő rendezett listájává alakít. Megfontoltan kell bánni ezzel az operátorral, hiszen a relációs algebra néhány művelete nincs értelmezve listára. Ezzel szemben a vetítés és a kiválasztás elvégezhető listákra is, sőt az eredményben a lista elemeinek sorrendje is megkövetelhető.

6. *Külső összekapcsolás*. Az összekapcsolásnak az egyik fajtája, amely a lógó sorokat is megőrzi. A lógó sorok `NULL` értékekkel „egészülnek ki" a külső összekapcsolás eredményében, tehát a lógó sorok is szerepelnek a végeredményben.

5.2.1. Ismétlődések megszüntetése

Néhány esetben szükségünk lehet olyan műveletre, amely a multihalmazból halmazt állít elő. A $\delta(R)$ kifejezést használjuk arra, hogy olyan halmazt kapjunk, amely R relációnak csak a különböző sorait tartalmazza.

5.8. példa. Tekintsük az 5.1. ábra R relációját:

A	B
1	2
3	4
1	2
1	2

Ekkor $\delta(R)$ a következő:

A	B
1	2
3	4

Figyeljük meg, hogy az $(1, 2)$ sor, amely $\delta(R)$-ben csak egyszer fordul elő, az R-ben háromszor is szerepelt. □

5.2.2. Összesítési műveletek

Több olyan művelet is létezik, amely alkalmazható számok vagy karakterláncok halmazaira vagy multihalmazaira. Ezek a műveletek összegzik vagy összesítik a reláció egy oszlopában szereplő értékeket. Ezen tulajdonságuk miatt ezeket összefoglalóan *összesítési műveletek*nek nevezzük. A legáltalánosabb műveletek ezek közül az alábbiak:

1. `SUM`, az oszlop értékeinek összegét határozza meg.
2. `AVG`, az oszlop értékeinek átlagát határozza meg.
3. `MIN`, illetve `MAX` az oszlop értékeinek minimumát, illetve maximumát határozza meg. Amennyiben karakterláncokat tartalmazó oszlopra alkalmazzuk, akkor a lexikografikusan (ábécé szerinti) legelső, illetve legutolsó értéket határozzák meg.
4. `COUNT`, az oszlopban található (nem feltétlenül különböző) elemek számát határozza meg. Vagy más szavakkal a `COUNT` egy reláció bármely attribútumára alkalmazva megadja a reláció sorainak a számát, beleértve az ismétlődéseket is.

5.9. példa. Tekintsük az alábbi relációt:

A	*B*
1	2
3	4
1	2
1	2

Tekintsünk néhány példát az adott reláció attribútumain történő összesítésekre:

1. `SUM(B)` $= 2 + 4 + 2 + 2 = 10$.
2. `AVG(A)` $= (1 + 3 + 1 + 1)/4 = 1.5$.
3. `MIN(A)` $= 1$.
4. `MAX(B)` $= 4$.
5. `COUNT(A)` $= 4$.

□

5.2.3. Csoportosítás

Néha nem egyszerűen egy oszlop összesítésére van szükségünk, hanem a reláció sorait kell csoportosítanunk egy vagy több oszlop értékei szerint, és így csoportonként összesíthetünk. Például, ha szeretnénk meghatározni mindegyik stúdió által előállított filmpercek összességét stúdiónként, azaz egy, az alábbihoz hasonló relációt:

stúdióNév	*hossz*
Disney	12345
MGM	54321
...	...

Induljunk ki a 2.2.8. alfejezetben szereplő adatbázissémából:

`Filmek(cím, év, hossz, műfaj, stúdióNév, producerAzon)`

Ezért csoportosítanunk kell a `Filmek` reláció sorait a `stúdióNév` szerint, majd csoportonként ki kell számolni a `hossz` összegét. Azaz tegyük fel, hogy a `Filmek` sorai az 5.4. ábrán szereplő alakban vannak, és csoportonként alkalmazzuk a `SUM(hossz)` összesítést.

	studioName	
	Disney Disney Disney	
	MGM MGM	
	○ ○ ○	

5.4. ábra. Egy reláció az elképzelt csoportfelosztásról

5.2.4. A csoportosítási művelet

Most már bevezethetünk egy olyan műveletet, ami lehetővé teszi egy reláció csoportokra osztását, illetve néhány oszlopra vonatkozó összesítést. Ha van csoportosítás, akkor az összesítés az egyes csoportokon belül értendő.

A γ alsó indexeként szereplő L elemeknek egy listája, ahol egy elem az alábbiak közül bármelyik lehet:

a) Az R reláció egy attribútuma, ahol γ R-re van alkalmazva. Azaz egyike R olyan attribútumainak, amelyre a csoportosítást végeztük. Ezt az elemet *csoportosítási attribútum*nak nevezzük.

b) A reláció valamelyik attribútumára vonatkozó összesítési művelet. Ha az összesítés eredményére névvel szeretnénk hivatkozni, akkor egy nyilat és egy új nevet kell utána írnunk. A kiindulásul szolgáló attribútumot *összesítési attribútum*nak nevezzük.

Az $\gamma_L(R)$ kifejezés eredményrelációjának felépítése a következő:

1. Osszuk R sorait *csoportok*ba. Egy csoport azokat a sorokat tartalmazza, amelyeknek az L listán szereplő csoportosítási attribútumokhoz tartozó értékei megegyeznek. Ha nincs csoportosítási attribútum, akkor az egész R reláció egy csoportot képez.

2. Minden csoporthoz hozzunk létre olyan sort, amelyik tartalmazza:

   i. A szóban forgó csoport csoportosítási attribútumait.

   ii. Az L lista összesítési attribútumaira vonatkozó összesítéseket. (Az adott csoport összes sorára.)

5.10. példa. Tegyük fel, hogy adott az alábbi reláció:

```
SzerepelBenne(cím, év, színészNév)
```

A δ speciális esete γ-nak

Technikailag a δ művelet redundáns. Ha az $R(A_1, A_2, \ldots, A_n)$ egy reláció, akkor $\delta(R)$ megegyezik a $\gamma_{A_1,A_2,\ldots,A_n}(R)$ kifejezéssel. Azaz az ismétlések megszüntetéséhez csoportosítást végzünk az attribútumokra, de nem összegezzük azokat. Ekkor minden csoportnak egy sor felel meg, amelyek R-ben egynél többször is előfordulhattak. Mivel a γ eredménye valójában csak egy sort tartalmaz a csoportokra, ezért a „csoportosításunk" eredményeként az ismétlődések is megszűnnek. A δ viszont egy elterjedt és fontos művelet, ezért célszerű továbbra is külön vizsgálnunk a műveletek algebrai szabályokkal és algoritmusokkal történő megvalósítása során.

Azt is láthatjuk, hogy a γ a halmazon alkalmazott vetítésművelet kiterjesztése. Azaz $\gamma_{A_1,A_2,\ldots,A_n}(R)$ eredménye ugyanaz, mint a $\pi_{A_1,A_2,\ldots,A_n}(R)$ kifejezésé, ha az R halmaz. Ha viszont az R multihalmaz, akkor a γ megszünteti az ismétlődéseket, míg a π nem.

Meg szeretnénk találni minden olyan színészt, aki már szerepelt legalább három filmben, illetve a hozzá tartozó első film dátumát is. Az első lépésben csoportosítanunk kell a `színészNév` mezővel, mint csoportosítási attribútummal. Mindegyik csoportra ki kell számolnunk a `MIN(év)` összesítést. Emellett viszont ki kell számítanunk a `COUNT(cím)` összesítést is minden csoportra ahhoz, hogy eldönthessük azt, mely csoport elégíti ki a legalább három filmben való szereplés feltételeit.

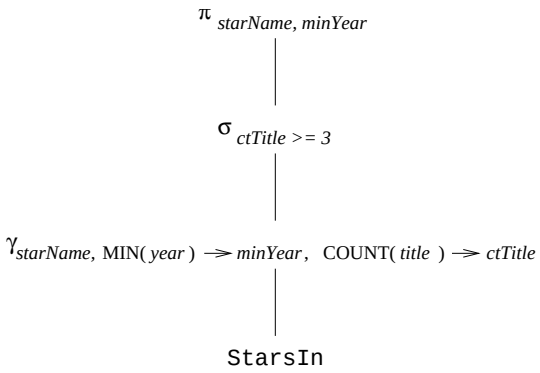
Elsőként írjuk fel a csoportosító kifejezést:

$$\gamma_{\texttt{színészNév},\ \texttt{MIN(év)}\rightarrow\texttt{elsőFilmÉv},\ \texttt{COUNT(cím)}\rightarrow\texttt{filmekSzáma}}(\texttt{SzerepelBenne})$$

A kifejezés első két oszlopa kelleni fog a végeredményben is. A harmadik oszlop egy segédattribútum (amelyet `filmekSzáma`-nak neveztünk) annak meghatározására, hogy az adott színész szerepelt-e már legalább három filmben. Azaz a lekérdezésnek megfelelő algebrai kifejezést egy `filmekSzáma >= 3` feltétellel történő kiválasztással és az első két oszlopra történő vetítéssel kell kiegészítenünk. A lekérdezés kifejezésfáját az 5.5. ábrán tekinthetjük meg. □

5.2.5. A vetítés művelet kiterjesztése

Tekintsük át újra a 2.4.5. részben bevezetett $\pi_L(R)$ vetítési műveletet. A klasszikus relációs algebrában az L az R reláció néhány attribútumának listájaként értelmezhető. Most pedig terjesszük ki a vetítés műveletét úgy, hogy lehetővé tegye számítások elvégzését is a sorok választott komponenseivel. A *kiterjesztett vetítés*, amelynek jelölése szintén $\pi_L(R)$, vetítési listájában a következő típusú elemek szerepelhetnek:



5.5. ábra. Az 5.10. példa lekérdezésének az algebrai kifejezésfája

1. R reláció egy attribútuma.
2. Egy $x \to y$ kifejezés, melyben x és y attribútum neveket jelölnek. Az L lista $x \to y$ eleme lekérdezi az R x attribútumát és y-ra *nevezi át* az oszlop nevét, azaz az eredmény sémájában ezen attribútum neve y lesz.
3. Egy $E \to z$ kifejezés, ahol E az R reláció attribútumaira vonatkozó (konstansokat, aritmetikai műveleteket, illetve karakterlánc-műveleteket tartalmazó) kifejezés, z pedig az E kifejezés alapján számolt, az eredményekhez tartozó új attribútumnak a nevét jelöli. Példaként tekintsük az $a + b \to x$ kifejezést a lista elemének, ekkor ez az a és b attribútumok összegét jelöli, amelynek a neve x lesz. A $c||d \to e$ elem jelentése pedig az, hogy c és d karakterláncként értelmezett attribútumokat összefűzzük, majd az eredményül kapott oszlopot e-nek nevezzük el.

A vetítés kiszámítása R összes sorának figyelembevételével történik. Az L lista kiértékelése a sor azon komponenseinek behelyettesítésével történik, amelyek szerepeltek L attribútumai között. A behelyettesítéskor az L-ben szereplő műveleteket is elvégezzük. Az eredmény egy olyan reláció lesz, amelynek a sémáját az L listában szereplő nevek (illetve átnevezések) alkotják. R minden egyes sora egy sort eredményez a végeredményben. Ezért az R relációban többször előforduló sorok az eredményben is többször fordulnak elő. A végeredményben viszont akkor is lehetnek ismétlődések, ha R-ben eredetileg nem voltak.

5.11. példa. Legyen az R reláció:

A	B	C
0	1	2
0	1	2
3	4	5

Ekkor $\pi_{A,B+C \to X}(R)$ értéke a következő:

A	X
0	3
0	3
3	9

Az eredmény sémájának két attribútuma van. Az első az A, ami R attribútuma volt ugyanezzel a névvel. A második az X, ami R második és harmadik attribútumának összegét reprezentálja.

Másik példaként tekinthetjük $\pi_{B-A\to X,C-B\to Y}(R)$ értékét, azaz:

X	Y
1	1
1	1
1	1

Figyeljük meg, hogy ennek a vetítéslistának a kiszámításánál a $(0,1,2)$ és a $(3,4,5)$ sorokat ugyanazon $(1,1)$ sorba képezzük le. Ezért ez a sor jelenik meg háromszor az eredményben. □

5.2.6. Rendezési művelet

Jó néhány esetben elképzelhető, hogy egy reláció sorait egy vagy több attribútuma alapján szeretnénk rendezni. Lekérdezéseink során gyakran megköveteljük az eredményreláció rendezettségét. Tegyük fel például, hogy Sean Connery filmjeit szeretnénk lekérdezni, és elvárjuk, hogy a kapott lista filmcím szerint rendezett legyen azért, hogy egy konkrét filmet gyorsabban megtalálhassunk. A lekérdezések optimalizálásának vizsgálatánál majd láthatjuk, hogyan válik gyakran hatékonyabbá az ABKR-ben egy lekérdezés végrehajtása, ha a relációt először rendezzük.

A $\tau_L(R)$ kifejezés, ahol R egy relációt, az L pedig az R reláció attribútumait tartalmazó listát jelenti, magát az R relációt határozza meg csak már az L lista alapján meghatározott rendezett alakban. Ha az L lista tartalma $A_1, A_2, \ldots, A_n$, akkor R sorait először A_1 értékei szerint fogja rendezni. A megmaradt csomókat A_2 értéke szerint tovább bontja, majd az olyan soroknál, amelyek mind A_1, mind A_2 szerint is azonos rendezettségűek, az A_3 értéke szerint rendezi, és így tovább. Azokat a csomókat, amelyek A_n értéke szerinti rendezés után is megmaradtak, tetszőleges sorrendben helyezzük el.

5.12. példa. Ha tekintjük az $R(A,B,C)$ sémájú R relációt, akkor a $\tau_{C,B}(R)$ az R sorait C érték szerint rendezi, majd az ugyanazon C értékkel rendelkező sorokat B szerint továbbrendezi. Ha mind a C, mind a B érték megegyezik, akkor tetszőleges lesz a sorrend. □

Amennyiben a τ által már rendezett eredményre egy olyan másik műveletet is elvégzünk, mint az összekapcsolás, akkor a korábbi rendezettség az esetek

nagy többségében jelentését veszti, és ezután az eredményre is inkább multihalmazként érdemes tekinteni, nem pedig listaként. Ezzel szemben a multihalmazokon értelmezett vetítések megőrzik a rendezettséget. Sőt, a listára alkalmazott kiválasztás eredményében, amely a kiválasztás feltételének nem megfelelő sorokat eldobja, a megmaradó sorok még mindig az eredeti rendezés szerinti sorrendjükben szerepelhetnek.

5.2.7. Külső összekapcsolások

Az összekapcsolás műveletének egyik tulajdonsága, hogy lehetnek lógó sorok, amelyek nem kapcsolhatóak össze más sorokkal, azaz nem lesz egyezés ezen sorok és a másik reláció sorai között a közös attribútumokon. A lógó sorokhoz nem tartozik sor az összekapcsolás eredményében, így az összekapcsolás nem biztos, hogy az eredeti reláció adatait hiánytalanul tükrözi. Az ilyen esetekben az összekapcsolás egy változatát, nevezetesen a „külső összekapcsolást", ajánlott alternatívaként használni. A külső összekapcsolás megtalálható a különféle, forgalomban lévő rendszerekben.

Először is tekintsük a „természetes" esetet, amelyben az összekapcsolás a benne szereplő két reláció közös attribútumaiban lévő értékeknek az egyenlőségén alapult. A $R \overset{\circ}{\bowtie} S$ alakú *külső összekapcsolás* elsőször elvégzi az $R \bowtie S$ természetes összekapcsolást, majd ehhez hozzáadja az R és az S lógó sorait is. Az előbbi módon hozzáadott sort minden olyan attribútumában ki kell egészíteni egy speciális *null* szimbólummal ($\perp$), amellyel a sor ugyan nem rendelkezett, de az összekapcsolás eredményében már szerepel. Megjegyezzük, hogy $\perp$ szimbólum a `NULL` (a 2.3.4. alfejezetben szereplő) értéknek felel meg SQL-ben.

5.13. példa. Vegyük az 5.6. (a) ábra, illetve az 5.6. (b) ábra U és V relációját. Az U $(1, 2, 3)$ sora így V-nek mind a $(2, 3, 10)$, mind a $(2, 3, 11)$ sorával összekapcsolható. Így ez a három sor nem lógó. Viszont a fennmaradó V-beli és U-beli sorok lógók: a $(4, 5, 6)$, illetve a $(7, 8, 9)$ az U-ból és a $(6, 7, 12)$ a V-ből. Azaz, ezen három sor egyikének sincs olyan sorpárja a másik relációban, amellyel a B és C komponenseken is megegyezne. Ezért az 5.6. (c) ábrán szereplő $U \overset{\circ}{\bowtie} V$ eredményében a lógó sorok ki vannak egészítve egy $\perp$ szimbólummal ott, ahol nem rendelkeznek értékkel: D attribútumnál az U sorainál és A attribútumnál a V sorai esetén. □

Az alap (természetes) külső összekapcsolás elvének létezik több variánsa is.

A $R \overset{\circ}{\bowtie}_L S$ *bal oldali külső összekapcsolás* annyiban különbözik a külső összekapcsolástól, hogy csak a bal oldalon szereplő R argumentum lógó sorainak $\perp$ szimbólummal kiegészített változatát adjuk hozzá az eredményhez.

A $R \overset{\circ}{\bowtie}_R S$ *jobb oldali külső összekapcsolás* annyiban különbözik a külső összekapcsolástól, hogy csak a jobb oldalon szereplő S argumentum lógó sorainak $\perp$ szimbólummal kiegészített változatát adjuk hozzá az eredményhez.

A	B	C
1	2	3
4	5	6
7	8	9

(a) U reláció

B	C	D
2	3	10
2	3	11
6	7	12

(b) V reláció

A	B	C	D
1	2	3	10
1	2	3	11
4	5	6	$\perp$
7	8	9	$\perp$
$\perp$	6	7	12

(c) $U \overset{\circ}{\bowtie} V$ eredmény reláció

5.6. ábra. Relációk külső összekapcsolása

5.14. példa. Ha vesszük az 5.6. ábra U és V relációját, akkor $U \overset{\circ}{\bowtie}_L V$:

A	B	C	D
1	2	3	10
1	2	3	11
4	5	6	$\perp$
7	8	9	$\perp$

Az $U \overset{\circ}{\bowtie}_R V$ pedig:

A	B	C	D
1	2	3	10
1	2	3	11
$\perp$	6	7	12

□

A három természetes összekapcsolási műveleten túl a théta-összekapcsolás még hátramaradt, amely során először egy théta-összekapcsolást végzünk, majd az eredményéhez hozzáadjuk azokat a $\perp$ szimbólummal kiegészített sorokat is, amelyeket a théta-összekapcsolás feltételének ellenőrzése során nem tudunk a másik reláció egyetlen sorával sem társítani. A $\overset{\circ}{\bowtie}_C$ kifejezést használjuk a C feltétellel rendelkező théta-összekapcsolás jelölésére. Ez a művelet is módosítható L vagy R segítségével bal vagy jobb oldali külső összekapcsolássá.

5.15. példa. Legyenek U és V az 5.6. ábra relációi, és tekintsük a következőt:

$$U \overset{\circ}{\bowtie}_{A>V.C} V$$

Az U $(4,5,6)$ és $(7,8,9)$ sorai, illetve a V $(2,3,10)$ és $(2,3,11)$ sorai is kielégítik a feltételt. Ezért ezen négy sor egyike sem lesz nem-társítható. Ezzel szemben a maradék két sor lógó lesz: az U $(1,2,3)$ sora és a V $(6,7,12)$ sora. Ezért kiegészítve fognak szerepelni az 5.7. ábrán szereplő eredményben. □

A	$U.B$	$U.C$	$V.B$	$V.C$	D
4	5	6	2	3	10
4	5	6	2	3	11
7	8	9	2	3	10
7	8	9	2	3	11
1	2	3	$\perp$	$\perp$	$\perp$
$\perp$	$\perp$	$\perp$	6	7	12

5.7. ábra. A théta-összekapcsolás eredménye

5.2.8. Feladatok

5.2.1. feladat. Tekintsük az alábbi két relációt:

$R(A,B)$: $\{(0,1),\ (2,3),\ (0,1),\ (2,4),\ (3,4)\}$

$S(B,C)$: $\{(0,1),\ (2,4),\ (2,5),\ (3,4),\ (0,2),\ (3,4)\}$

Számítsuk ki a következőket: a) $\pi_{A+B,A^2,B^2}(R)$; b) $\pi_{B+1,C-1}(S)$; c) $\tau_{B,A}(R)$; d) $\tau_{B,C}(S)$; e) $\delta(R)$; f) $\delta(S)$; g) $\gamma_{A,\ \mathtt{SUM}(B)}(R)$; h) $\gamma_{B,\mathtt{AVG}(C)}(S)$; ! i) $\gamma_A(R)$; ! j) $\gamma_{A,\mathtt{MAX}(C)}(R \bowtie S)$; k) $R \overset{\circ}{\bowtie}_L S$; l) $R \overset{\circ}{\bowtie}_R S$; m) $R \overset{\circ}{\bowtie} S$; n) $R \overset{\circ}{\bowtie}_{R.B<S.B} S$.

! 5.2.2. feladat. Egy f egyértékű műveletet *idempotens*nek nevezünk, ha bármely R relációra teljesül, hogy $f\big(f(R)\big) = f(R)$, azaz f többszöri alkalmazása ugyanazt az eredményt adja, mintha egyszer alkalmaztuk volna. A következő operátorok közül melyik lesz idempotens? Válaszához adjon számolási példát vagy indokolja meg.

a) δ; b) π_L; c) σ_C; d) γ_L; e) τ.