Jelölések áttekintése

Elevenítsük fel a 6.2.3. részben a relációk méretének jelölésére használt konvenciókat:

- $B(R)$ jelöli az R reláció összes sorának tárolásához szükséges blokkok számát.
- $T(R)$ az R reláció sorainak számát jelöli.
- $V(R, a)$ az R reláció a attribútumához tartozó *értékszámlálót* jelenti, vagyis azoknak a különböző értékeknek a számát, amelyek az R relációban az a attribútum értékeként előfordulnak. Valamint $V(R,[a_1, a_2, \ldots, a_n])$ jelöli azoknak a különböző értékeknek (értékkombinációknak) a számát, amelyek előfordulnak R-ben, amikor az $a_1, a_2, \ldots, a_n$ attribútumokat együtt tekintjük, azaz a $\pi_{a_1, a_2, \ldots, a_n}(R)$-ben szereplő különböző sorok számát jelenti.

7.4.2. Vetítés méretének becslése

A vetítés abban különbözik a többi művelettől, hogy az eredményének a mérete kiszámítható. Mivel egy vetítés minden argumentumsorhoz előállít egy eredménysort, a kimenet méretének változása csak a sorok hosszának megváltozásában jelentkezik. Emlékezzünk, hogy az itt használt vetítés operátor egy multihalmaz operátor, és nem távolítja el az ismétlődéseket. Ha egy vetítés során előálló ismétlődéseket meg akarjuk szüntetni, akkor a δ operátort kell utána alkalmazni.

Normális esetben vetítéskor a sorok összezsugorodnak, hiszen bizonyos komponenseket elhagyunk. A vetítésnek a 6.1.3. részben bevezetett általános formája azonban megengedi új komponensek létrehozását, mint attribútumok kombinációit. Vannak tehát esetek, amikor egy π operátor növeli a reláció méretét.

7.4.3. Kiválasztás méretének becslése

Amikor egy kiválasztást hajtunk végre, általában csökkentjük a sorok számát, de a sorok mérete ugyanaz marad. A kiválasztás legegyszerűbb esetében, amikor egy attribútumnak egy konstanssal való egyenlőségét vizsgáljuk, létezik egy könnyű módszer az eredmény méretének becslésére, feltéve, hogy tudjuk (vagy becsülni tudjuk) az attribútum által felvett különböző értékek számát. Legyen $S = \sigma_{A = c}(R)$, ahol A az R egy attribútuma és c egy konstans. Ekkor a következő becslést javasoljuk:

- $T(S) = T(R)/V(R, A)$

Ez a szabály biztosan igaz akkor, ha az A atribútum minden értéke egyenlő gyakorisággal fordul elő az adatbázisban. A fenti szabály azonban még akkor is a legjobb becslése az átlagnak, ha az A értékei nem mutatnak egyenletes eloszlást az adatbázisban. Elvárjuk viszont, hogy az A minden értéke egyforma valószínűséggel szerepeljen az A értékét meghatározó lekérdezésekben.

Problematikusabb a méret becslése, amikor a kiválasztás egyenlőtlenség-összehasonlítást tartalmaz, például ha $S = \sigma_{a < 10}(R)$. Azt gondolhatnánk, hogy az átlag tekintetében a sorok fele megfelelne az összehasonlításnak, a sorok fele nem, így $T(R)/2$ jó becslése lenne az S méretének. Egy érzés azonban azt súgja, hogy egy ilyen lekérdezés a lehetséges soroknak inkább csak egy kisebb hányadát adná vissza.[1] Egy olyan szabályt javasolunk, amely figyelembe veszi ezt a tendenciát, és azzal a feltételezéssel él, hogy egy tipikus vizsgálat, amely az egyenlőtlenséget vizsgálja, körülbelül a sorok egyharmadát adja vissza, nem a felét. Ha $S = \sigma_{a < c}(R)$, akkor $T(S)$-re a becslésünk:

- $T(S) = T(R)/3$

A „nem egyenlő” összehasonlítások ritkák. Ha azonban egy olyan kiválasztással találkozunk, mint például az $S=\sigma_{a \neq 10}(R)$, akkor javasoljuk annak feltételezését, hogy lényegében minden sor kielégíti majd ezt a feltétel.

[1] Ha például fizetésekről lennének adataink, azt kérdeznénk-e meg nagyobb valószínűséggel, hogy a fizetés *kisebb,* mint 500 000 Ft, vagy azt, hogy *nagyobb,* mint 500 000 Ft?

Vehetjük tehát becslésként a következőt: $T(S) = T(R)$. Egy másik becslés lehet a $T(S) = T(R)(V(R, a)–1)/V(R, a)$, ami valamivel kevesebbet ad. Ez a megközelítés elismeri, hogy az R sorainak körülbelül $1/V(R, a)$ része elbukik a feltételen, mert azok a értéke egyenlő a konstanssal.

Amikor egy C kiválasztási feltétel több ÉS-sel összekötött egyenlőségvizsgálat vagy más összehasonlítás, akkor a $\sigma_C(R)$ kiválasztást úgy tekinthetjük, mint azoknak az egyszerű kiválasztásoknak egymás utáni alkalmazását, amelyek mindegyike a feltétel egy-egy részét ellenőrzi. Vegyük észre, hogy ezeknek a kiválasztásoknak a sorrendje nem számít. Ennek hatásaként az eredmény méretére vonatkozó becslés az lesz, hogy az eredeti reláció méretét megszorozzuk az egyes feltételekhez tartozó *szelektivitási* tényezőkkel. Ez a tényező 1/3 egyenlőtlenség esetén, 1 ≠ esetén, illetve $1/V(R, A)$ amikor a C feltételben egy A attribútumot hasonlítunk egy konstanshoz.

Amikor egy kiválasztás VAGY-gyal kapcsolt feltételeket tartalmaz, mondjuk $S = \sigma_{C_1 \text{ OR } C_2}(R)$, kevesebb bizonyosságunk van az eredmény méretét illetően. Egy egyszerű feltételezés az, hogy egyetlen sorra sem teljesül mindkét feltétel, vagyis az eredmény mérete egyenlő az egyes feltételeket kielégítő sorok számának összegével. Ez a becslés általában túlbecslést jelent, és néha valóban ahhoz az abszurd következtetéshez vezethet el minket, hogy az S-ben több sor van, mint az eredeti R relációban. Egy másik egyszerű megközelítés lehet, hogy vesszük a minimumát az R méretének, és annak, amit a C_1-et, illetve a C_2-t kielégítő sorok számának összegeként kapunk.

Egy kevésbé egyszerű, de feltehetően pontosabb becslést kapunk az

$$S = \sigma_{C_1 \text{ OR } C_2}(R)$$

méretére, ha feltesszük, hogy C_1 és C_2 függetlenek. Ekkor, ha R-nek n sora van, amelyek közül m_1-re teljesül a C_1, és m_2-re teljesül a C_2, akkor az S-ben megjelenő sorok számára a következő becslést adhatjuk:

$$n\big(1 - (1 - m_1/n)(1 - m_2/n)\big)$$

Itt az $1 – m_1/n$ egyenlő a soroknak a C_1-et nem teljesítő a hányadával, $1 – m_2/n$ pedig a soroknak a C_2-et nem teljesítő a hányadát jelenti. Ezek szorzata a R sorainak azon hányadát adja, amelyek *nincsenek* benne az S-ben, és ezt a szorzatotot 1-ből kivonva az S-ben szereplő hányadot kapjuk.

7.24. példa: Tegyük fel, hogy az $R(a, b)$ relációnak $T(R) = 10000$ sora van, és legyen

$$S = \sigma_{a=10 \text{ OR } b < 20}(R)$$

Legyen $V(R, a) = 50$. Ekkor az $a = 10$ feltételt kielégítő sorok számát, ami $T(R)/V(R, a)$, 200-ra becsüljük. A $b < 20$ feltételt kielégítő sorok számát $T(R)/3$-ra, vagyis 3333-ra becsüljük.

Az S méretére vonatkozó legegyszerűbb becslés ezek összege, azaz 3533. Az $a = 10$ és $b < 20$ feltételek függetlenségére építő bonyolultabb becslés a

$$10\,000\big(1 - (1 - 200/10\,000)(1 - 3333/10\,000)\big)$$

értéket adja, azaz 3466-ot. A két becslés között kicsi az eltérés, így nagyon valószínűtlen, hogy az egyik választása a másikkal szemben változást jelentene a legjobb fizikai terv kiválasztásában. □

Az utolsó operátor, amely egy kiválasztási feltételben szerepelhet: a NOT. Ha egy R relációnak n számú sora van, akkor a NOT C feltételt kielégítő sorok becsült számát úgy kapjuk meg, hogy n-ből kivonjuk a C-t kielégítő sorok becsült számát.

7.4.4. Összekapcsolás méretének becslése

Csak a természetes összekapcsolást fogjuk vizsgálni. A többi összekapcsolás az alábbi elveknek megfelelően kezelhető:

1. Egy egyenlőség alapú összekapcsolás (equijoin) eredményében megjelenő sorok száma, miután a változó nevekben bekövetkező változásokkal elszámoltunk, pontosan úgy számítható ki, mint természetes összekapcsolás esetén. Ezt a pontot a 7.26. példa fogja szemléltetni.
2. Más théta-összekapcsolások úgy becsülhetők, mintha szorzatot követő kiválasztások volnának, figyelembe véve a következő további megjegyzéseket:

a) Egy szorzat sorainak számát úgy kapjuk, hogy a szorzatban részt vevő relációk sorainak számait összeszorozzuk.
b) Egy egyenlőséget vizsgáló összehasonlítást a természetes összekapcsoláshoz kidolgozott technika segítségével becsülhetünk.
c) Egy két attribútum egyenlőtlenségét vizsgáló $R.a < S.b$ típusú összehasonlítás úgy kezelhető, mint egy $R.a < 10$ alakú összehasonlítás, amit a 7.4.3. részben tárgyaltunk. Vagyis feltehetjük, hogy ennek a feltételnek a szelektivitási tényezője 1/3 (ha úgy gondoljuk, hogy a feltétel inkább ritkán teljesül), vagy lehet 1/2 (ha nem élünk a feltételezéssel).

Első körben tételezzük fel, hogy két reláció természetes összekapcsolása csak két attribútum egyenlőségét tartalmazza. Ez azt jelenti, hogy az $R(X, Y) |><| S(Y, Z)$ összekapcsolást vizsgáljuk, de kezdetben feltesszük, hogy Y egyetlen attribútum, az X és Z viszont tetszőleges attribútum halmazokat jelölhetnek.

Az a probléma, hogy nem tudjuk, hogy az R és S Y értékei milyen viszonyban állnak egymással. Például:

1. A két relációban az Y értékek lehetnek diszjunkt halmazok, amikor is az összekapcsolás üres és $T(R |><| S) = 0$.
2. Az Y lehet az S kulcsa és egy idegen kulcs az R-ben. Ilyenkor az R minden egyes sora pontosan egy S-beli sorral kapcsolódik, így tehát $T(R |><| S) = T(R)$.
3. Lehet, hogy az R és S majdnem minden sorának ugyanaz az Y értéke, ekkor $T(R |><| S)$ körülbelül $T(R)T(S)$ lesz.

A következő két egyszerűsítő feltételezéssel fogunk élni, hogy a leggyakoribb esetekre koncentrálhassunk:

1. *Értekhalmazok tartalmazása.* Ha Y egy több relációban is szereplő attribútum, akkor ez az attribútum mindegyik relációban egy $y_1, y_2, y_3, \ldots$ rögzített értéklistának az elejéről kap értéket, és az összes érték ebből a prefixből származik. Következésképpen, ha R és S két reláció, amelyek tartalmazzák az Y attribútumot, és $V(R, Y) \leq \leq V(S, Y)$, akkor az R minden Y értéke az S-nek Y értéke lesz.
2. *Értékhalmazok megőrzése*. Ha egy R relációt összekapcsolunk egy másik relációval, akkor egy A attribútum, amely nem összekapcsolási attribútum (azaz nem szerepel mindkét relációban), nem veszít el értékeket az értékeinek a lehetséges halmazából. Pontosabban szólva, ha A az R-nek attribútuma, de S-nek nem, akkor $V(R |><| S, A) = V(R, A)$. Megjegyezzük, hogy az R és az S összekapcsolásának sorrendje nem lényeges, tehát azt is mondhattuk volna, hogy $V(S |><| R, A) = V(R, A)$.

Nyilván előfordulhat, hogy az 1. előfeltevés, értékhalmazok tartalmazása, nem érvényes, de teljesül akkor, ha Y kulcs az S-ben, és idegen kulcs az R-ben. Sok más esetben is megközelítőleg igaz, hiszen intuitíve azt várjuk, hogy ha S-nek sok Y értéke van, akkor egy adott R-ben előforduló Y érték jó eséllyel szerepel S-ben.

A 2. feltételezés, értékhalmazok megőrzése, szintén sérülhet, de igaz a feltevés akkor, ha az $R |><| S$ összekapcsolási attribútuma kulcs az S-ben, és idegen kulcs az R-ben. Valójában csak akkor fordulhat elő, hogy a 2. előfeltevés nem teljesül, ha az R-ben „lógó sorok” vannak, vagyis olyan sorok, amelyek az S egyetlen sorával sem kapcsolódnak, de még az ilyen esetekben is érvényes lehet az előfeltétel.

E feltételezések mellett az $R(X, Y) |><| S(Y, Z)$ mérete a következőképpen becsülhető. Legyen $V(R, Y) \leq V(S, Y)$. Ekkor $1/V(S, Y)$ az esélye annak, hogy az R egy t sora az S egy adott sorával kapcsolódik. Mivel az S-nek $T(S)$ sora van, azoknak a soroknak a várható száma, amelyekkel t kapcsolódik: $T(S)/V(S, Y)$. Minthogy az R-nek $T(R)$ sora van, az $R |><| S$ becsült mérete $T(R)T(S)/V(S, Y)$. Ha $V(R, Y) \geq V(S, Y)$, akkor a szimmetria alapján kapott becslés: $T(R |><| S) = T(R)T(S)/V(R, Y)$. Általában a $V(R, Y)$ és a $V(S, Y)$ közül a nagyobbal osztunk, tehát:

- $T(R |><| S) = T(R)T(S)/\max\big(V(R, Y), V(S, Y)\big)$

7.4.7. Egyéb műveletek méretének becslése

Láttunk két műveletet, ahol az eredményül kapott sorok száma egy pontos fomulával leírható:

1. A vetítés nem változtatja meg egy relációban szereplő sorok számát.
2. A szorzat olyan eredményt állít elő, amelyben a sorok száma egyenlő az argumentum relációkban lévő sorok számának szorzatával.

Két további műveletre – a kiválasztásra és az összekapcsolásra – elég jó becslési technikákat dolgoztunk ki. A fennmaradó műveletek esetében azonban nem könnyű az eredmény méretének meghatározása. Sorra vesszük a többi relációs algebrai operátort is, és javaslatokat fogunk tenni arra, hogy ez a becslés hogyan végezhető el.

Egyesítés

Ha a multihalmaz-egyesítést vesszük, akkor a méret pontosan az argumentumok méretének összegével egyenlő. Egy halmazegyesítésnél a méret lehet olyan nagy, mint a **méretek összege**, **vagy** olyan kicsi, mint a két argumentum mérete közül **a nagyobb**. Azt ajánljuk, hogy válasszunk valamit a kettő között félúton, például az összeg és a nagyobb átlagát (ami ugyanaz, mint a nagyobb plusz a kisebb fele).

Metszet

Az eredménynek lehet olyan kevés sora, mint például **0, vagy** olyan sok sora, mint a két argumentum közül a **kisebb**nek, függetlenül attól, hogy halmaz- vagy multihalmaz-metszetről van szó. Egy lehetséges megközelítés, hogy a szélsőségek közti átlagot vesszük, ami a kisebb felét jelenti.

Egy másik lehetőség, hogy felismerjük azt, hogy a metszet a természetes összekapcsolás egy speciális esete, és a 7.4.4. részben bevezetett formulát használjuk. Halmazmetszet esetén ez a formula garantáltan olyan eredményt ad, ami nem nagyobb, mint a két reláció közül a kisebb. Egy multihalmaz-metszet esetében azonban előfordulhatnak rendellenességek, amikor a becslés nagyobb, mint bármelyik argumentum. Nézzük például az $R(a, b) \cap_M S(a, b)$ metszetet, ahol az R a (0, 1) sor két példányából áll, és az S ugyanennek a sornak három példányából áll. Ekkor $V(R, a) = V(S, a) = V(R, b) = V(S, b) = 1$, $T(R) = 2$ és $T(S) = 3$. Az összekapcsolásra vonatkozó szabály alapján a becslés $2 \times 3/\big(\max(1, 1) \times \max(1, 1)\big) = 6$, de az eredményben nyilvánvalóan nem lehet több, mint $\min\big(T(R), T(S)\big) = 2$ sor.

Különbség

Amikor az $R - S$ különbséget vesszük, akkor az eredményben megkapott sorok száma ***T*(*R*) és *T*(*R*) – *T*(*S*) között** lehet. Becslésként az átlagot javasoljuk: $T(R) - T(S)/2$.

Ismétlődések megszüntetése

Ha $R(a_1, a_2, \ldots, a_n)$ egy reláció, akkor a $\delta(R)$ mérete $V(R,[a_1, a_2, \ldots, a_n])$. Sokszor azonban nem rendelkezünk ezzel a statisztikai értékkel, ezért közelíteni kell. Mint szélsőségek, a $\delta(R)$ mérete **megegyezhet az *R* méretével** (nincsenek ismétlődések, **vagy lehet 1** (az R minden sora ugyanaz).[2] Egy másik felső korlát a $\delta(R)$-ben levő sorok számára az elképzelhető különböző sorok száma: a $V(R, a_i)$-k szorzata, ahol $i = 1, 2, \ldots, n$. Ez a szám lehet kisebb, mint a $T\big(\delta(R)\big)$ más becslései. Több olyan szabály is van, amit használhatnánk a $T\big(\delta(R)\big)$ becslésére. Az egyik elfogadható az, hogy a $T(R)/2$ és az összes $V(R, a_i)$ szorzata közül vesszük a kisebbiket.

Csoportosítás és összesítés

Tegyük fel, hogy van egy $\gamma_L(R)$ kifejezésünk, és e kifejezés eredményének méretére kell becslést adnunk. Ha rendelkezünk a $V(R,[g_1, g_2, \ldots, g_k])$ statisztikával, ahol a g_i-k az L-ben szereplő csoportosítási attribútumok, akkor az lesz a válaszunk. A statisztika azonban esetleg nem elérhető, így szükségünk van egy másik módszerre, amivel a $\gamma_L(R)$ méretét becsülhetjük. A $\gamma_L(R)$ sorainak száma megegyezik a csoportok számával. Az eredményben lehet egy csoport, vagy lehet olyan sok csoport, mint ahány sor van az R-ben. A δ-hoz hasonlóan, a csoportok számára a $V(R, A)$-k szorzatával is adhatunk felső korlátot, de itt az A attribútum csak az L csoportosítási attribútumain fut végig. Újfent azt a becslést javasoljuk, ami veszi a $T(R)/2$ és e szorzat közül a kisebbiket.

[2] Szigorúan véve, ha R üres, akkor sem az R-ben, sem a $\delta(R)$-ben nincs sor, vagyis az alsó korlát 0. Csakhogy ritkán érdekel bennünket ez a speciális eset.