Adatbázisok elmélete 14. előadás

Katona Gyula Y.
Budapesti Műszaki és Gazdaságtudományi Egyetem
Számítástudományi Tsz.
I. B. 137/b
`kiskat@cs.bme.hu`
`http://www.cs.bme.hu/~kiskat`

2004

Normalizálás

Tétel. *Tetszőleges (R, F) sémának van hűséges felbontása BCNF relációkra.*

Bizonyítás: Elve:

- Ha (R, F) BCNF $\surd$
- Ha nem, akkor két valódi (kisebb) részre bontjuk hűségesen $\Longrightarrow (R_1, R_2)$
- Ezt ismételjük (R_1, R_2)-re.

Ez véget fog érni, mert ha már csak 2 attribútum marad valamelyikben, azt nem kell tovább bontani.

Hűséges lesz, mert láttuk, hogy ha egy hűséges felbontás egyik részét tovább bontjuk, akkor hűséges marad.

Bizonyítás

Hogyan bontjuk fel 2 valódi részre, hűségesen?

Keresünk a felbontandó sémában egy olyan $X \to A \in F^+$-t, ami sérti a BCNF tulajdonságot
$\Longrightarrow$ A és X része a sémának, $A \notin X$ és X nem szuperkulcs

$$R_1 := XA, \qquad R_2 := R \setminus \{A\}$$

Ezek kisebbek: R_2 nyilván, R_1 pedig azért, mert ha $R_1 = R$ volna, akkor $X \to XA = R$ miatt X szuperkulcs lett volna.

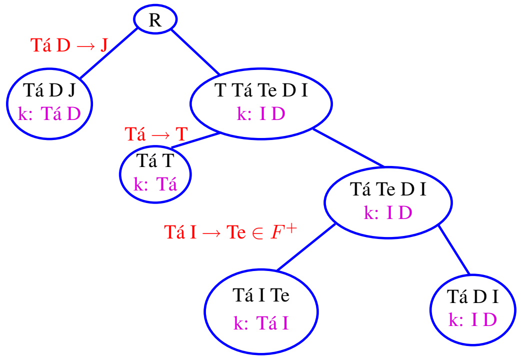
Hűséges a felbontás: kétrészes teszttel $R_1 \cap R_2 = X \to A = R_1 \setminus R_2$ $\surd$

Miért lesz jobb ez a felbontás?
Az $X \to A$ függéssel nem lesz több probléma: R_2-ben nincs A, így nem lehet baj. R_1-ben viszont X szuperkulcs lesz.

Példa

R(**T**anár, **Tá**rgy, **Te**rem, **D**iák, **J**egy, **I**dő)

F = {Tá → T; IT → Te; ID → Te; ID → Tá; TáD → J} $\Longrightarrow$ kulcs csak ID



Megjegyzések

Minden felbontás után meg kell nézni, hogy a kapott relációk BCNF-ben vannak-e. Ehhez meg kell konstruálni F_S^+-et, ha S a vizsgált reláció: ez az F_R^+ azon függéseiből áll, amiknek mindkét oldala F-ben van. Ezeket a függéseket úgy kapjuk, hogy minden $X \subseteq S$ részhalmazra kiszámoljuk $X^+(F)$-et és $X \to Y$ pontosan akkor lesz benne F_S^+-ben, ha $Y \subseteq X^+(F) \cap S$.

Általában nem igaz, hogy elég F-ből kiválogatni azokat, amiknek mindkét oldala S-ben van.

Pl.: $F = \{$Tá $\to$ T; IT $\to$ Te; ID $\to$ Te; ID $\to$ Tá; TáD $\to$ J$\}$
Ha $S =$ Tá Te D I, akkor (csak a nemtrivi függéseket felírva):
$F_S^+ = \{$Tá I $\to$ Te; D I $\to$ Te Tá; D I Tá $\to$ Te; D I Te $\to$ Tá$\}$

$\Longrightarrow$ Az előző algoritmus lehet exponenciális $\Longrightarrow$ Van polinomiális algoritmus is.

3 attribútum esetén a BCNF tulajdonság csak úgy sérülhet, ha $X \to Y$, ahol X, Y egy-egy attribútum és X nem kulcs.

Azt is mindig ellenőrizni kell, hogy a kapott relációkban mik a (szuper)kulcsok, hogy egy függésről el tudjuk dönteni, hogy sérti-e a BCNF-et vagy nem. A példában ez viszonylag könnyű lesz, hiszen I és D egyik F-beli függőségben sem szerepel a jobb oldalon, így minden kulcs (amikor I és D szerepel a relációban) tartalmazza I D-t.
Csak azt kell megnézni, hogy I D kulcs marad.

Függőség megőrzése

BCNF egy fogyatékossága: nehéz lehet ellenőrizni, hogy teljesülnek-e F függései (pl. beszúráskor). Ilyenkor a költséges $\bowtie$ kell, és ez sokszor előfordulhat.
Kéne egy olyan felbontás, amin könnyen lehet ellenőrizni a függéseket.

Definíció. *Adott (R, F) séma és ennek egy $\rho = (R_1, \dots, R_k)$ felbontása.*

$$\pi_\rho(F) := \{X \to Y \in F^+ \mid \exists i\ (1 \le i \le k)\ X, Y \subseteq R_i\}^+$$

*az F **függéseinek vetítése a ρ felbontásra**. ρ **függőségőrző**, ha $\pi_\rho(F) = F^+$.*

Megjegyzés: $\pi_\rho(F) \subseteq F^+$ persze mindig igaz.

Ha a felbontás függőségőrző, akkor elég a darabokon ellenőrzni valamit, ami garantálja, hogy F minden függése fennmarad az egészen.

Példa

R(**V**áros, **U**tca, **I**rányítószám) $F = \{VU \to I;\ I \to V\}$
Ez nem BCNF $I \to V$ miatt.

Mire jó a függőségőrzés?:
Ha felbontjuk $\Longrightarrow S(V, I),\ Q(I, U)$
Beszúrunk 2-2 sort:

S	V	I
	Nagykanizsa	8800
	Nagykanizsa	8831

Q	U	I
	Kossuth	8800
	Kossuth	8831

Noha S-ben és Q-ban oké minden, $S \bowtie Q$-ban nem teljesül a $VU \to I$ függés.
Ez nem lett volna, ha függőségőrző lenne a felbontás.

Szomorú példa ez: semelyik felbontása sem őrzi meg $VU \to I$-t, mert csak ez olyan függés, aminek jobb oldalán van I, azaz ha egy felbontás függőségőrző lenne, akkor egy tagjában kéne VUI-nek lennie, de az nem lenne valódi felbontás.
$\Longrightarrow$ ennek nincs függőségőrző valódi felbontása, vagyis van olyan reláció, amit nem lehet függőségőrző módon BCNF-ekre szétszedni

Következmény

Állítás. *Felbontás BCNF-be nem feltétlenül függőségőrző.*

Kellene egy gyengébb normálforma. Ebben lehet valamennyi redundancia, de legyen függőségőrző.

3NF

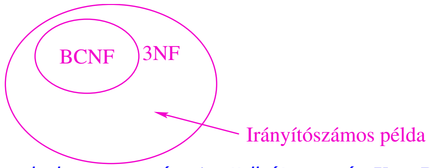
Definíció. *Az (R, F) séma A attribútuma* ***prím*** *(elsődleges), ha szerepel valamelyik kulcsban.*

Szuperkulcsban minden szerepel, kulcs helyett szuperkulccsal nem lenne sok értelme az előbbi definíciónak.

Definíció. *Az (R, F) séma* ***3NF*** *(harmadik normálformájú), ha tetszőleges nemtriviális $X \to A \in F^+$ függés esetén vagy X szuperkulcs vagy A prímattribútum.*

Következmény. *Minden BCNF séma egyben 3NF is.*

3NF lehet redundáns, de nem nagyon.



Tétel. *Ha (R, F) egy 3NF séma, akkor minden nem prím A attribútumra és $X \subseteq R$ kulcsra igaz, hogy nincs olyan Y, hogy $X \to Y$, $Y \not\to X$, $Y \to A$ és $A \notin Y$.* *(Nem-elsődleges attribútum nem függ tranzitívan kulcstól.)*

Nem bizonyítjuk, úgy menne, mint BCNF-nél a hasonló állítás.

3NF tulajdonság ellenőrzése

Hogyan tudjuk ellenőrizni, hogy egy séma 3NF-ben van-e?

Állítás. *≤ 2 attribútumos reláció mindig 3NF.*

Bizonyítás: Már láttuk, hogy BCNF $\Longrightarrow$ 3NF

Mivel F^+-ra követeljük meg a feltételt, nehéz ellenőrizni.
DE:

Tétel. *Ha (R, F) nem 3NF, akkor van olyan $X \to Y \in \mathrm{F}$, amely jobboldalának valamely A attribútumára $X \to A$ nemtriviális, X nem szuperkulcs és A nem prím.* *(Az ilyen $X \to A \in F^+$.)*

Nem bizonyítjuk.

Azt tudjuk ellenőrizni egy adott $X \to A$ függésre, hogy X szuperkulcs-e: kiszámítjuk $X^+(F)$-et.

De hogyan ellenőrizzük, hogy A prím-e? $\Longrightarrow$ kell az **összes** kulcs

Tétel. *Annak eldöntése, hogy egy attribútum prím-e, NP-teljes probléma.*

Következmény. *Annak eldöntése, hogy egy séma 3NF-ben van-e, NP-teljes probléma.*

3NF tulajdonság ellenőrzése

Persze olyan algoritmus van, ami legrosszabb esetben exponenciális:

- Meghatározzuk az összes kulcsot.
- Meghatározzuk az összes prímattribútumot.
- Minden F-beli $X \to Y$ függésre nézzük meg:
  - ★ Igaz-e, hogy $Y \subseteq X$. Ha igen, a függés triviális. $\surd$
  - ★ Igaz-e, hogy X kulcs-e. Ha igen, nem sérti a feltételt. $\surd$
  - ★ Igaz-e, hogy Y-ben csak prímattribútumok vannak. Ha igen, nem sérti a feltételt. $\surd$

  Ha egyik sem, akkor van olyan függés, ami sérti a feltételt $\Longrightarrow$ nem 3NF

3NF felbontás

Tétel. *Tetszőleges (R, F) sémának van hűséges és függőségőrző felbontása 3NF sémákra.*

Definíció. *A G függéshalmaz az F függéshalmaz* ***fedése*** *ha $G^+ = F^+$.* *(Persze ilyenkor F is fedése G-nek.)*

Definíció. *A G függéshalmaz az F függéshalmaz* ***minimális fedése****, ha egyrészt fedése, másrészt*

(1) a G-beli függések $X \to A$ alakúak, ahol $A \notin X$

(2) G-ből nem hagyható el függés: $(G \setminus \{X \to A\})^+ \subsetneq G^+$

(3) G-beli függések baloldalai minimálisak: $Y \subsetneq X \Longrightarrow (G \setminus \{X \to A\} \cup \{Y \to A\})^+ \subsetneq G^+$

Állítás. *Tetszőleges F-nek van minimális fedése.*

Bizonyítás: Algoritmust adunk, külön gondoskodunk minden pont teljesítéséről.

(1) $X \to Y \in G$, $Y = A_1 \dots A_k$ $\Longrightarrow$ minden $X \to A_i$-t beveszünk, ha $A_i \notin X$.

(2) Minden $X \to A \in G$ függésre kiszámoljuk $Y := X^+(G \setminus \{X \to A\})$-t. Ha $A \in Y$, akkor $X \to A$ elhagyható, különben nem.

(3) Ellenőrizni kell, hogy $X \to A$ baloldala minimális-e. X minden B elemére kiszámoljuk $Y := (X \setminus \{B\})^+(G)$-t. Ha $A \in Y$, akkor $X \to A$ helyett vegyük be $X - \{B\} \to A$-t. Ha egyik B-re se lesz ilyen, akkor X minimális.

Megjegyzés: És persze a fenti három lépés során a függéshalmaz lezártja nem változik.

Példa

$R = (A, B, C, D)$ $\quad F = \{AB \to CD;\ AC \to BD;\ C \to A;\ C \to B\}$

(1) $F' = \{AB \to C;\ AB \to D;\ AC \to B;\ AC \to D;\ C \to A;\ C \to B\}$

(2) $C \to B$ miatt $AC \to B$ elhagyható és $AB \to C$ és $AC \to D$ miatt $AB \to D$ elhagyható, de más nem, ezt végig lehet nézni. $F' = \{AB \to C;\ AC \to D;\ C \to A;\ C \to B\}$

(3) $C \to A$ miatt $AC \to D$ baloldaláról A elhagyható.
$F'' = \{AB \to C;\ C \to D;\ C \to A;\ C \to B\}$

Ez már minimális fedés.

A minimális fedés nem feltétlenül egyértelmű!

Példa: $R(A, B, C)$ $\quad F = \{AB \to C;\ A \to B;\ B \to A\}$ esetén jó minimális fedés lesz
$G_1 = \{B \to C;\ A \to B;\ B \to A\}$ és
$G_2 = \{A \to C;\ A \to B;\ B \to A\}$ is.

Bizonyítás

Tétel. *Tetszőleges (R, F) sémának van hűséges és függőségőrző felbontása 3NF sémákra.*

Bizonyítás: Vegyük F egy minimális fedését: $G = \{X_1 \to A_1, \ldots, X_k \to A_k\}$
Legyen X egy kulcs és $\rho = (X, X_1A_1, \ldots, X_kA_k)$ egy felbontás.
$\implies R_0, R_1, \ldots, R_k$
Állítás: ez függőségőrző lesz, a tagok 3NF-ek és a felbontás hűséges.

ρ függőségőrző: $F^+ = G^+$ és minden G-beli $X \to A_i$ függés benne lesz R_i-ben (ott ellenőrizhető).

R_0 3NF: R_0-ban nincs nemtriviális függés, mert különben X nem lenne kulcs, csak szuperkulcs $\implies R_0$ BCNF $\implies$ 3NF

Többi R_i is 3NF: tegyük fel, hogy nem az $\implies \exists\, U \to B$ nemtriviális függés, hogy U nem szuperkulcs R_i-ben és B nem prímattribútum R_i-ben.
Ha $B = A_i$, akkor $U \subseteq X_i$, de $U \neq X_i$, hiszen akkor U szuperkulcs lenne R_i-ben. $\implies U \subset X_i$
$\implies X_i \to A_i$ baloldala csökkenthető G-ben U-ra. Ellentmondás, mert akkor G nem volt minimális fedés.
Ha $B \neq A_i$ $\implies B \in X_i$ és B nem prím R_i-ben $\implies X_i$ nem kulcs R_i-ben (de szuperkulcs)
$\implies \exists\, Y \subsetneq X_i$ kulcs R_i-ben $\implies Y \to A_i$ fennáll $\implies X_i \to A_i$ baloldala csökkenthető G-ben Y-ra, megint csak ellentmondás.

ρ hűséges: Higgyük el, nem bizonyítjuk.

Megjegyzés: Előfordulhat, hogy valamelyik X_iA_i már tartalmaz kulcsot. Ilyenkor a $\rho = \{X_1A_1, \ldots, X_kA_k\}$ is jó felbontás már.

Megjegyzés: 2NF már nem érdekes, 1NF kicsit érdekes, de nem foglalkozunk vele.

Példa: 3NF-re bontás

$R = (A, B, C, D, E)$ $\quad F = \{AE \to BC;\ AC \to D;\ CD \to BE;\ D \to E\}$

Ez nem 3NF, mert a kulcsok:
semelyik egyelemű halmaz nem kulcs (csak D lehetne, de az ő lezártja csak DE), viszont kételeműek közül szuperkulcs lesz AC, AD, AE (A-nak benne kell lennie minden kulcsban, mert A nincs jobboldalon), AB viszont nem szuperkulcs.
Ezek kulcsok is lesznek, mert egyik egyelemű se volt kulcs.
Más kulcs nincs is, mert ha lenne legalább háromelemű halmaz, aminek a lezártja az egész, akkor abban A biztos benne van és legalább C vagy D vagy E is benne van, de akkor az már csak szuperkulcs lehet, mert tartalmaz kulcsot.

Innen látszik, hogy a prímattribútumok: A, C, D, E, vagyis B nem az.

Példa: 3NF-re bontás (folyt.)

$R = (A, B, C, D, E)$ $F = \{AE \to BC;\ AC \to D;\ CD \to BE;\ D \to E\}$

Tehát a $CD \to B$ függés rossz a 3NF szempontjából, mert CD nem szuperkulcs és B nem prím.

Csináljunk hát egy 3NF-ekre való függőségőrző, hűséges felbontást.

(1) $F' = \{AE \to B;\ AE \to C;\ AC \to D;\ CD \to B;\ CD \to E;\ D \to E\}$

(2) $AE \to C, AC \to D, CD \to B$ miatt $AE \to B$ elhagyható és $D \to E$ miatt $CD \to E$ is elhagyható, de más nem, ezt végig lehet nézni (mert például AE-nek a maradék függésekre vett lezártjában nincsen benne C).
$F'' = \{AE \to C;\ AC \to D;\ CD \to B;\ D \to E\}$

(3) Semelyik baloldal nem csökkenthető, mert például A lezártjában nincsen benne E, és a többi is ugyanígy látszik. Vagyis F'' már minimális fedés.

A minimális fedés alapján a jó felbontás:
(AEC, ACD, CDB, DE) mivel kulcsot nem is kellett hozzávennünk, mert az már benne van az egyik tagban (pl. AE az elsőben).